AD/A-005 891

THE MONTE CARLO METHOD OF EVALUATING INTEGRALS

Daniel T. Gillespie

Naval Weapons Center
China Lake, California

February 1975

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS
BEFORE COMPLETING FORM

1. REPORT NUMBER
NWC TP 5714

2. GOVT ACCESSION NO.

3. RECIPIENT'S CATALOG NUMBER
AD/A-005891

4. TITLE (and Subtitle)
THE MONTE CARLO METHOD OF EVALUATING INTEGRALS

5. TYPE OF REPORT & PERIOD COVERED

6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s)
Daniel T. Gillespie

8. CONTRACT OR GRANT NUMBER(s)

9. PERFORMING ORGANIZATION NAME AND ADDRESS
Naval Weapons Center
China Lake, CA 93555

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
Task Area No. ZF000-01-01

11. CONTROLLING OFFICE NAME AND ADDRESS

12. REPORT DATE
February 1975

13. NUMBER OF PAGES
210

14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)

15. SECURITY CLASS. (of this report)
UNCLASSIFIED

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)
Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)
Monte Carlo
Numerical Integration
Random Numbers

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)
See back of form.

Reproduced by
NATIONAL TECHNICAL
INFORMATION SERVICE
US Department of Commerce
Springfield, VA. 22151

(U) The Monte Carlo Method of Evaluating Integrals, by Daniel T. Gillespie. China Lake, California, Naval Weapons Center, February 1975. 203 pp. (NWC TP 5714, publication UNCLASSIFIED.)

(U) This monograph is a tutorial presentation of the Monte Carlo method of numerically estimating definite integrals. Intended primarily for scientists and engineers, and assuming very little background in probability theory, the monograph attempts to convey a clear understanding of what the Monte Carlo method is, why it works, and how it can be used to evaluate complicated (and often otherwise intractable) multidimensional integrals. General methods are developed for generating sets of random points having prescribed biases; procedures are outlined for using such random points to obtain an estimate of the value of a given integral and the uncertainty associated with this estimate; and techniques are described for reducing the uncertainty without significantly increasing the computation time. Background material and peripherally related topics are discussed in an extensive set of appendices.

CONTENTS

Appendices

Figures

Chapter 1

INTRODUCTION

The evaluation of one or more specific definite integrals often constitutes a crucial segment of a scientific or engineering research project. As technical research becomes more complex, the integrals encountered tend to be correspondingly more complicated, and so more resistant to treatment by standard mathematical methods. In particular, scientists and engineers are increasingly being confronted with the task of evaluating multi-dimensional integrals, an enterprise which often overtaxes even highly refined, computer-oriented numerical quadrature techniques.

In the late 1940's a novel technique for numerically evaluating integrals was suggested by E. Fermi, J. von Neumann and S. Ulam. This technique, which was termed "Monte Carlo" because of its reliance upon random numbers, did not win immediate widespread acceptance in the scientific community. Most scientists apparently just ignored Monte Carlo, probably because it seemed rather foolish to suppose that one could gain useful knowledge about a well-defined and perfectly deterministic integral by playing some contrived game of chance. And of those workers who took the trouble to inform themselves more fully on the subject, many found that Monte Carlo, as it was understood and applied in the 1950's, simply was not as efficient for their particular problems as were the more conventional numerical methods.

To be sure, the Monte Carlo method has its limitations, and it is by no means an appropriate tool for many problems. However, owing to an increased understanding of and improvement in Monte Carlo techniques, as well as the development of faster digital computers, the class of integrals which are now amenable to Monte Carlo is fairly large, and includes many of the unwieldy multi-dimensional integrals which scientists and engineers often encounter in their research. For this reason the Monte Carlo approach undoubtedly deserves a wider recognition in the scientific and engineering community than it presently has.

There exists a fairly extensive body of literature on Monte Carlo. Currently the most comprehensive work is a book by Hammersley and Handscomb (Ref. 1); a less ambitious but somewhat more readable work is an article by Fluendy (Ref. 2). However, in this writer's opinion the "standard work" on Monte Carlo has yet to be written. This is probably because all of its variations and possibilities have not yet been brought into a completely understood and totally unified picture by any one practioner of the art. In addition, it is usually easier to do Monte Carlo in some specific instance than it is to write (or read) about it in general terms. Unfortunately though, it is also easy to do Monte Carlo inefficiently—or worse still, incorrectly—so a good understanding of the generalities is rather essential.

This writer has used Monte Carlo as a computational tool in two areas of physics, namely, elementary particle physics (Ref. 3) and classical kinetic theory (Ref. 4). This limited experience has by no means rendered the author an expert in all facets of Monte Carlo; however, it has suggested a pedagogical approach to the subject which is perhaps more transparent to scientists than are the standard presentations, which usually tend to be rather deeply couched in the technical language of statistics. The aim of this monograph, therefore, is not to provide a definitive and comprehensive treatment of all aspects of Monte Carlo; rather, its purpose is to present the basic principles of the conventional Monte Carlo method for estimating integrals, in a manner that will convey an "intuitive feeling" for how and why the method works. An intuitive rapport with the Monte Carlo approach is important, because this enables one to more easily identify which features of a given integral will give rise to difficulties, as well as which features can be exploited for a gain in computing efficiency. More often than not, this kind of insight is what spells the difference between success and failure in obtaining a sufficiently accurate numerical estimate for a given integral.

From the viewpoint of a scientist, the basic idea behind Monte Carlo can probably be best explained through a familiar example from statistical mechanics: Suppose we have a gas composed of very many molecules of mass m in thermal equilibrium at absolute temperature T. If f(v) is any function of the molecular speed v (e.g., f(v) could be the molecular kinetic energy $mv^2/2$, or the molecular speed v itself), then the average $\langle f \rangle$ of f(v) for these gas molecules may be defined as

$$\langle f \rangle = \frac{1}{N} \sum_{i=1}^{N} f(v_i), \qquad N>>1 \tag{1.1}$$

where $v_1, v_2, \ldots, v_N$ are the speeds of N *randomly chosen* molecules. Now, we can evaluate $\langle f \rangle$ without actually polling N randomly selected molecules by making use of the Maxwell-Boltzmann Law. According to this law, the *probability* that any molecule will have a speed between v and v+dv is

$$P(v)dv = (\alpha/\pi)^{3/2}\ 4\pi v^2 \exp(-\alpha v^2)dv \tag{1.2}$$

where $\alpha \equiv m/2kT$, k being Boltzmann's constant. It follows that the contribution to the sum on the right of (1.1) coming from molecules with speeds between v and v+dv will be NP(v)dv×f(v). Summing (integrating) over all dv-intervals thus gives the quantity $\Sigma_i f(v_i)$, and (1.1) yields

$$\langle f \rangle = \int_0^\infty f(v)P(v)dv \tag{1.3}$$

The point here is that, in statistical mechanics, we evaluate *averages* of the kind on the right side of (1.1) by actually computing *definite integrals* of the kind on the right side of (1.3).

The basic idea behind Monte Carlo is simply to turn this procedure around. Thus, suppose that for some unrelated reason we wanted to evaluate the integral on the right side of (1.3), where P(v) is given by (1.2) with a specific numerical value for α, and where f(v) is some given function which is so complicated that we are unable to carry out the integration analytically. Now, if we could somehow obtain a set of numbers $v_1, v_2, \ldots, v_N$ that mimic the speeds of N randomly chosen gas molecules in thermal equilibrium (with m and T values appropriate to the given value of α), then we could evidently evaluate the integral in question simply by averaging f(v) over these N v_i-values. This would in fact constitute a "Monte Carlo evaluation" of the integral: In Monte Carlo, we evaluate definite integrals of the kind on the right side of (1.3) by actually computing averages of the kind on the right side of (1.1).

From the foregoing rough description of the Monte Carlo approach, several questions naturally arise. The first and most obvious is, can we obtain the required v_i-values without actually measuring the speeds of randomly selected gas molecules? More generally, can we obtain v_i-values appropriate to P functions different from the one given in (1.2)? These matters will be addressed in Chapter 2, where we discuss in some detail how sets of random points are specified and how they can be constructed.

The second question concerns the accuracy of a Monte Carlo calculation. If the values $v_1, v_2, \ldots, v_N$ are the speeds of N randomly chosen gas molecules, then we surely cannot expect (1.1) to yield a unique result. In fact, (1.1) - (1.3) are strictly valid only in the limit $N \to \infty$. We therefore need to know what sort of uncertainty in our result will be occasioned by calculating (1.1) with N finite. This question will be discussed in detail in Chapter 3, where we shall develop the Monte Carlo method for estimating definite integrals in a much more careful way than we did above.

Finally, there is the following question: Once the uncertainty in a given Monte Carlo calculation has been determined, is there any way of modifying the procedure so as to reduce this uncertainty? One intuitively obvious way of doing this would be to simply increase N, but clearly the time available for computation will impose an effective upper limit on the size of N. It turns out that, depending on the specifics of the integral in question, one usually can find ways of reducing the uncertainty without significantly increasing the computing time. In Chapter 4 we shall describe some of these so-called "variance reducing" techniques.

We mentioned that this report will concentrate on the "conventional" Monte Carlo method of evaluating integrals. As implied, there is a somewhat unconventional Monte Carlo method; this alternate approach makes use of the mathematical concept of a "Markov chain" or a "Markovian random walk", and has met with considerable success in certain areas of statistical mechanics. It is not our purpose in this report to discuss in detail the Markov chain Monte Carlo method for calculating multi-dimensional integrals; however, in order to give the reader some idea of what is involved, as well as some guidance to the literature, we have included a brief appendix on this subject (Appendix I) at the end of this report.

We shall try in this report to avoid as much as possible the technical jargon of statistics, but we shall nevertheless attempt to maintain a reasonable level of precision and rigor. We assume at the outset that the reader is acquainted with the common (albeit not universal) view of "probability" as the ratio of the number of trials with a favorable outcome to the total number of trials, taken in the limit of infinitely many trials. From this notion one easily deduces the addition and multiplication laws for probabilities:

- Addition Law: If p_1 and p_2 are the probabilities for the occurrence of two mutually exclusive events 1 and 2, then the probability for the occurrence of either 1 or 2 in any one trial is p_1+p_2.

- Multiplication Law: If p_1 is the probability for the occurrence of event 1, and p_{21} the probability for the occurrence of event 2 when event 1 occurred on the previous trial, then the probability for the occurrence of first 1 and then 2 in any two successive trials is $p_1 \cdot p_{21}$.

These and other primitive notions about probabilities will be invoked frequently throughout our discussion of the Monte Carlo procedure.

Chapter 2

SETS OF RANDOM POINTS

2-1. Specifying Sets of Random Numbers

All Monte Carlo applications involve the use of at least one set of random numbers $\{x_i\}$ distributed according to some predetermined probability density function $P(x)$. By these terms we mean an inexhaustible set of real numbers from which we may "draw" sequential elements $x_1, x_2, x_3, \ldots,$ such that

$$P(x)dx \equiv \text{probability that any } x_i \text{ will lie between } x \text{ and } x{+}dx. \tag{2.1}$$

The numbers in $\{x_i\}$ are considered "random" because each draw can produce any real number x, provided $P(x)\neq 0$, and it is not possible to say beforehand what the drawn number will be. However, to say that the numbers in $\{x_i\}$ are "random" is not to say that they are "unbiased". Indeed, the numbers are quite definitely biased in the sense that, in the limit of infinitely many draws, a normalized frequency histogram of the x_i's will coincide with the curve $P(x)$-versus-x.

A set of random numbers $\{x_i\}$ is specified as completely as is possible by its probability density function $P(x)$. However, it is often convenient to work with its probability distribution function $F(x)$, which is defined in terms of $P(x)$ by

$$F(x) \equiv \int_{-\infty}^{x} P(x')dx' \tag{2.2}$$

In light of (2.1), (2.2) says that F(x) is the "sum" of the probabilities for x_i to fall inside each infinitesimal interval between $-\infty$ and x; by the addition law of probabilities, F(x) may thus be interpreted as *the probability that x_i will be less than x*.

Since x_i will surely be less than ∞, we have the following *normalization property*:

$$\int_{-\infty}^{\infty} P(x')dx' \equiv F(\infty) = 1 \tag{2.3}$$

Another property of P(x), which derives directly from its definition (2.1), is that it never be negative:

$$P(x) \geqslant 0 \quad \text{for all } x \tag{2.4}$$

It follows from (2.2)-(2.4) that the distribution function F(x) rises from the value 0 at $x=-\infty$ to the value 1 at $x=+\infty$ in a *non-decreasing* way.

Indeed, any non-negative, single-valued function of x which bounds a unit area with the x-axis can serve as a probability *density* function, defining a set of random numbers. Similarly, any differentiable function of x which rises from the value 0 at $x=-\infty$ to the value 1 at $x=+\infty$ without ever decreasing can serve as a probability *distribution* function, defining a set of random numbers.

The distinction between a probability density function and a probability distribution function is quite important in Monte Carlo work.[†]

[†] The function P(v) in (1.2), or its closely related Cartesian counterpart, is usually referred to as the "Maxwell-Boltzmann distribution function"; this is a rather unfortunate designation since it is obviously a probability *density* function and not a probability distribution function.

F(x) is the integral curve of P(x), and conversely P(x) is the derivative curve of F(x). A plot of P(x) and F(x) for a hypothetical set of random numbers $\{x_i\}$ is shown in Fig. 1, where we have tried to illustrate the properties and relationships developed above. If P(x) is zero below x=a and above x=b, then F(x) is zero below x=a and unity above x=b. The total area under the P(x) curve is unity, while the area under the P(x) curve between x and x+dx is numerically equal to the probability that a number drawn from $\{x_i\}$ will lie between x and x+dx. The ordinate at x on the F(x) plot is the probability that a number drawn from $\{x_i\}$ will be less than x. Regions on the x-axis of high likelihood are distinguished by high P(x)-values and steeply rising F(x)-values; regions of low likelihood are distinguished by low P(x)-values and nearly constant F(x)-values. Since F(x) is a probability, it is always a pure number between 0 and 1. P(x) is not a probability; however, P(x)dx is, so P(x) always has dimensions of 1/x.

F(x) is sometimes referred to as the "cumulative distribution function". We shall hereafter refer to P(x) and F(x) more simply as the "density function" and "distribution function" respectively.

Suppose a given set of random numbers $\{x_i\}$ with density function $P_1(x)$ is transformed into a new set of random numbers $\{y_i\}$ by applying to each element of $\{x_i\}$ the transformation

$$y = f(x) \qquad (2.5a)$$

What will be the density function $P_2(y)$ of the new set $\{y_i\}$? If, as is indicated

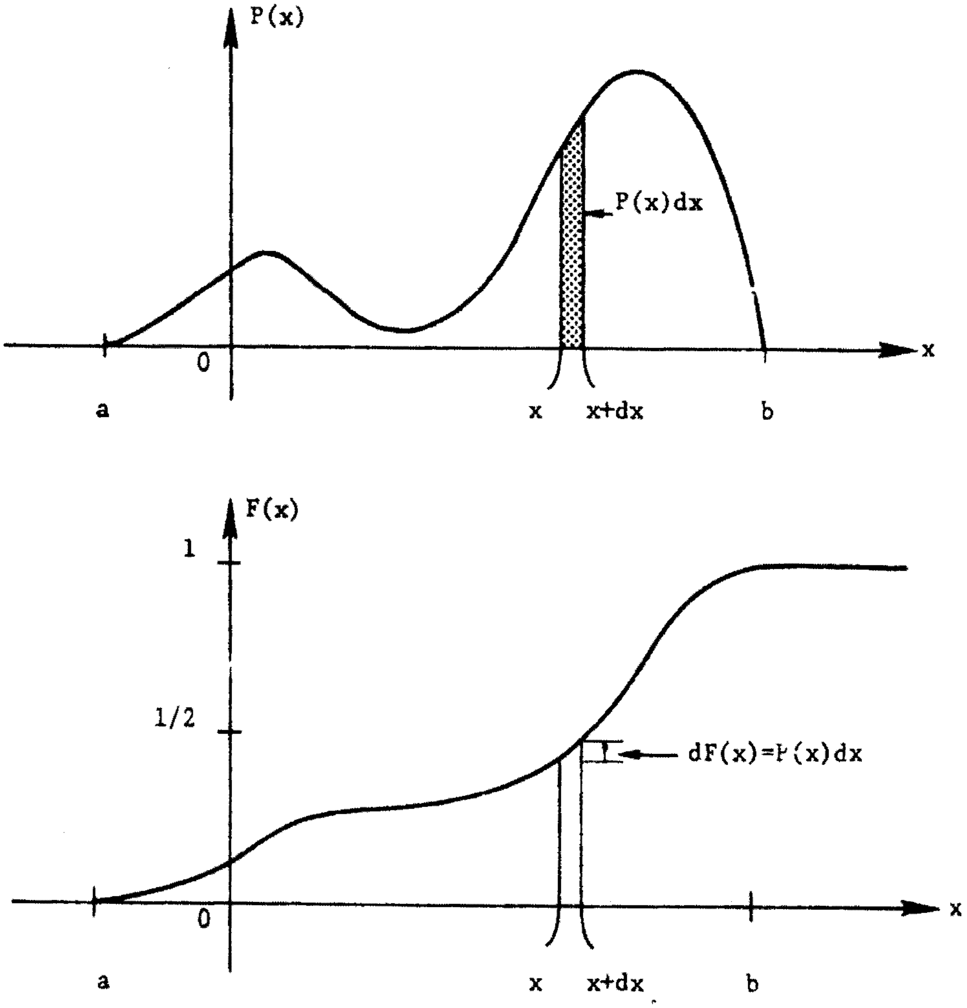

FIGURE 1. Illustrating the relationship between the density function P(x) and the distribution function F(x) for a hypothetical set of random numbers $\{x_i\}$.

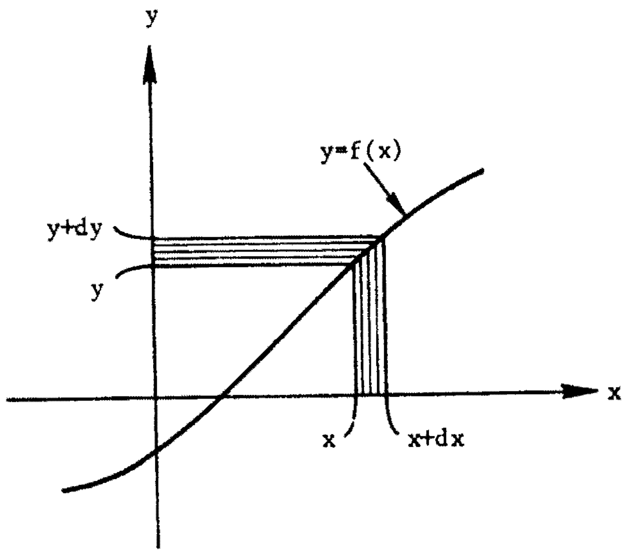

FIGURE 2. Transforming a set of random numbers $\{x_i\}$ into a set of random numbers $\{y_i\}$ through a function y=f(x).

in Fig. 2, the interval (y,y+dy) is the image of the interval (x,x+dx) under (2.5a), so that

$$dy = \left|\frac{dy}{dx}\right| dx = \left|f'(x)\right| dx \tag{2.5b}$$

then clearly the probability for finding y_i inside (y,y+dy) is the same as the probability for finding x_i inside (x,x+dx):

$$P_2(y)dy = P_1(x)dx \tag{2.6}$$

Inserting (2.5b) we therefore conclude that

$$P_2(y) = P_1(x)\Big/\left|f'(x)\right| \tag{2.7}$$

where x on the right side of (2.7) is <u>now</u> to be regarded as a function of y through the inverse of (2.5a): $x=f^{-1}(y)$. The important result (2.7) shows that the density of random points y_i around $y=f(x)$ will be greater than, equal to, or less than the density of random points x_i around x accordingly as the local slope $|dy/dx|$ of the transformation curve is less than, equal to, or greater than unity; these features can be appreciated geometrically from Fig. 2. If the inverse function $x=f^{-1}(y)$ is multivalued, so that a given dy-interval is populated from several dx intervals, then the right sides of (2.6) and (2.7) will evidently have to be summed over all contributing intervals.

2-2. The Set $\{r_i\}$

Of special importance in Monte Carlo work is the set of random numbers distributed uniformly over the unit interval, the elements of which we shall always denote by r_i. More precisely, the set $\{r_i\}$ is defined by the density function

$$P(r) = \begin{cases} 0, & \text{for } r<0 \\ 1, & \text{for } 0 \leq r \leq 1 \\ 0, & \text{for } r>1 \end{cases} \tag{2.8a}$$

or the corresponding distribution function [cf. (2.2)]

$$F(r) = \begin{cases} 0, & \text{for } r<0 \\ r, & \text{for } 0 \leq r \leq 1 \\ 1, & \text{for } r>1 \end{cases} \tag{2.8b}$$

Thus, the set $\{r_i\}$ is distinguished by the facts that: (i) the probability for a randomly drawn r_i to lie in any dr-interval between 0 and 1 is equal to dr; and (ii) the probability for a randomly drawn r_i to be less than a given number r between 0 and 1 is equal to r.

The set $\{r_i\}$ is important in Monte Carlo work for two reasons: First, there exist many short computer subroutines which are capable of rapidly generating elements of this set (or more precisely, elements of some set which simulates $\{r_i\}$ closely enough for most practical purposes); and second, it is possible to construct from the elements of the set $\{r_i\}$ the elements of a set $\{x_i\}$ distributed according to any prescribed density function P(x). In this report we shall not delve into the first point in any detail. The reason for this omission is that the writing of computer codes to generate mock elements of the set $\{r_i\}$—so-called "uniform random number generators"—is a complicated, fast-changing art which is best entrusted to experts in statistics and the theory of numbers. A nice introduction to this subject containing many references to the literature is the short article by Chambers (Ref. 5); more detailed treatments may be found in Chapter 3 of Hammersley and Handscomb (Ref. 1) and Vol. 2 of Knuth (Ref. 6). We shall content ourselves here with giving only a brief glimpse of the general ideas involved in generating uniformly distributed "pseudorandom" numbers on a digital computer.

Most uniform random number generators currently in use are based upon the so-called "multiplicative congruential method". In its simplest form, this method takes a starting integer N_o and generates a sequence of integers N_1, N_2, ..., by means of the recursion relation

$$N_i \equiv CN_{i-1} \text{ (modulo M)}$$

where C and M are predetermined (and usually very large) integers. This relation means that N_i is set equal to the remainder obtained when CN_{i-1} is divided by M. Obviously, each N_i will lie between 0 and M, so the elements

$$r_i = N_i/M$$

will lie between 0 and 1. It turns out that, provided sufficient care is taken in choosing the numbers N_o, C and M, the set of numbers $\{r_i\}$ obtained from the above algorithm approximates a uniform distribution of random numbers in the unit interval surprisingly well. What undesirable correlations the method has (and it certainly does have some) can be greatly diminished by incorporating a few twists and turns into the above procedure. However, almost all uniform random number generating subroutines generally available for digital computers have in common with the procedure just described the feature that each element r_k is calculated in an operationally simple way from the result of the r_{k-1} calculation, and also the feature that r_1 is determined by a starter number N_o whose value can be changed at will by the user to generate different, independent "chains" of random numbers. Usually it is most economical to set up a uniform random number generating subroutine so that, after an "initializing call" which sets some value for the starter number N_o, the subroutine will calculate and output one random number (the next number of the chain) each time it is called by the main Monte Carlo program.

The author's recent Monte Carlo work has made use of a short Fortran subroutine designed especially for the Univac 1108 computer by Marsaglia and Bray (Ref. 7); their method essentially tries to overcome some of the correlations present in congruential generators by mixing several such generators together. We refer the reader to their article and to the previously mentioned works (Refs. 5, 1 and 6) for further details on the computer-generation of pseudorandom numbers from a uniform distribution in the unit interval. In the sequel we shall simply assume that we have easy computer access to a set of numbers which effectively mimics the set $\{r_i\}$; in practice, this is usually the case.

2-3. The Inversion and Rejection Generating Methods

We turn now to the important problem of how to construct, from a *given* set of random numbers $\{r_i\}$ distributed uniformly on the unit interval, another set $\{x_i\}$ distributed according to any prescribed density function P(x). There are two primary methods for accomplishing this, which we shall refer to as the *inversion method* and the *rejection method*. We consider first the

> *Inversion Method*: Determine the distribution function F(x) corresponding to the given density function P(x) [cf. (2.2)]. Then, for each element r_i from the given set $\{r_i\}$, choose x_i by solving the equation $F(x_i)=r_i$; i.e., construct the elements of the set $\{x_i\}$ from the elements of the set $\{r_i\}$ according to the formula
>
> $$x_i = F^{-1}(r_i) \qquad (2.9)$$
>
> where F^{-1} is the inverse of the distribution function.

That the set $\{x_i\}$ constructed according to the foregoing procedure actually has P(x) as its density function follows from the transformation theorem proved at the end of Sec. 2-1 [cf. (2.5)-(2.7)]. Thus, if the set $\{r_i\}$ with density function $P_1(r)$ is transformed into a new set $\{x_i\}$ by the transformation $x=F^{-1}(r)$, then by (2.7) the density function $P_2(x)$ of the new set is

$$P_2(x) = P_1(r)/|F^{-1'}(r)|$$

But the density function of the set $\{r_i\}$ is just $P_1(r)=1$ $(0 \leq r \leq 1)$; furthermore, since $(dx/dr)=1/(dr/dx)$, then $(F^{-1})'=1/(F')$. Thus, the density function of the constructed set $\{x_i\}$ is

$$P_2(x) = 1/[1/F'(x)] = F'(x) \equiv P(x)$$

where the concluding equality follows from the definition (2.2).

To get some physical insight into the way the inversion method actually works, consider the hypothetical plot of $r=F(x)$-versus-x, shown in Fig. 3. Essentially, the inversion method lays out the elements of the given set $\{r_i\}$ along the r-axis, and then projects each r_i-element onto the x-axis through the curve $r=F(x)$. The projection is always well-defined since $F(x)$ rises from 0 at $x=-\infty$ to 1 at $x=+\infty$ in a non-decreasing way. If Δr_1 and Δr_2 are two equal-size intervals in $0<r<1$, then they will each contain the same number of elements of the set $\{r_i\}$, at least to within random statistical fluctuations, since the numbers in $\{r_i\}$ are uniformly distributed over the unit interval. By construction, then, the respective image intervals Δx_1 and Δx_2 will also contain the same number of elements of the set $\{x_i\}$, again to within random statistical fluctuations. Now if, as is the case in Fig. 3, the <u>slope</u> of the curve $F(x)$-versus-x is greater in Δx_2 than in Δx_1, then Δx_2 will be proportionately smaller than Δx_1, implying that Δx_2 will have a proportionately greater <u>density</u> of points than Δx_1. But the local slope of the curve $F(x)$-versus-x is just the local value of $P(x)$, as is seen from the definition (2.2). Thus, we see that the density of

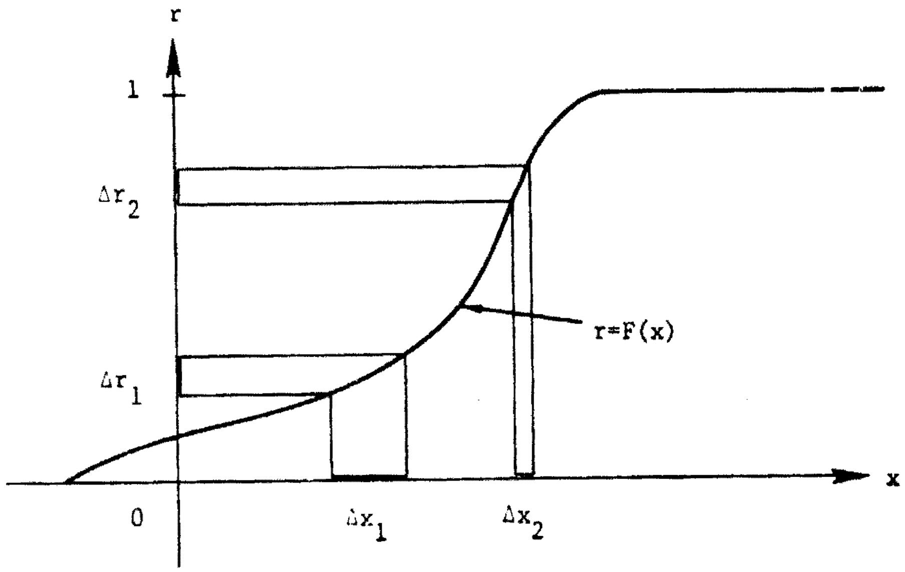

FIGURE 3. Illustrating the principle of the inversion generating method.

x_i-points produced in a given region by the inversion generating method is proportional to the value of the function P(x) in that region, which is just as it should be.

A simple but often used application of the inversion method is the generation of a set of random numbers $\{x_i\}$ distributed uniformly over the interval $a \leqslant x \leqslant b$. The density function here is evidently

$$P(x) = \begin{cases} 1/(b-a), & a \leqslant x \leqslant b \\ 0 \quad , & \text{otherwise} \end{cases} \tag{2.10a}$$

Using (2.2) we find that the distribution function in the interval $a \leqslant x \leqslant b$ is

$$F(x) = (x-a)/(b-a) \tag{2.10b}$$

The inversion of F(x) here is easily accomplished, and the construction rule (2.9) takes the entirely plausible form

$$x_i = a + r_i(b-a) \tag{2.10c}$$

As a second general procedure for generating random numbers $\{x_i\}$ according to a prescribed density function P(x), we consider the

Rejection Method: For this method it is required that the given density function P(x) vanish everywhere outside some finite interval $a \leqslant x \leqslant b$, and be bounded by some finite number B inside that interval. Furthermore, in addition to the set of random numbers $\{r_i\}$ distributed uniformly over the unit interval, we shall also need an independent set of random numbers $\{x_i'\}$ distributed uniformly over the interval $a \leqslant x \leqslant b$ [see (2.10)]. The generating procedure is then as follows. Draw a pair of random numbers (x_i', r_i) from the given sets, and take x_i' to be a member of the set $\{x_i\}$ if

$$P(x_i')/B \geqslant r_i \qquad (2.11)$$

If (2.11) is not satisfied, reject the pair (x_i', r_i) and keep drawing new pairs until the inequality is satisfied.

The proof that the set of x_i'-values which pass the "acceptance criterion" (2.11) i⁻ indeed distributed according to the density function P(x) is somewhat more complicated than the proot for the inversion method, and is presented in Appendix A. We merely point out here that the acceptance criterion (2.11) is evidently statistically favorable to x_i'-values for which P(x) is relatively large, and is statistically unfavorable to x_i'-values for which P(x) is relatively small. We should also note that, because a ratio is taken in (2.11), P(x) and its upper bound B need be known only up to an overall constant factor; i.e., the "normalization constant" need not be known when using the rejection method. In any case, one finds that the efficiency of this generating process, or the probable fraction of the x_i'-values which will be accepted as x_i-values, is given by [see Appendix A]

$$E = \frac{\int_a^b P(x)dx}{B \cdot (b-a)} \qquad (2.12)$$

Geometrically, E can be interpreted as the ratio of the area under the curve $P(x)$ to the area under the rectangle of height B and width $(b-a)$ which encloses the $P(x)$ curve. Clearly, then, it is desirable to choose for B the *smallest* upper bound on $P(x)$, and for (a,b) the *smallest* interval outside of which $P(x)$ vanishes identically.

In our derivation of the rejection method in Appendix A, it will be seen that this method can actually be formulated in a slightly more general way: If the initial set of random numbers $\{x_i'\}$ is distributed over $a \leq x \leq b$ according to some density function $\tilde{P}(x)$ [*not* necessarily the uniform function in (2.10a)], then the density function of the set $\{x_i\}$ which is constructed according to the selection process (2.11) will be $C\tilde{P}(x)P(x)$, C being the appropriate normalizing constant. This way of generating according to a product density function is usually less efficient than an "all-at-once" approach, but in some situations it may prove to be more convenient.

To compare in a few words the inversion and rejection methods for generating random numbers $\{x_i\}$ according to a prescribed density function $P(x)$, we may say that the inversion method constructs the set $\{x_i\}$ by distorting a uniformly distributed set through the *distribution* function, while the rejection method constructs the set $\{x_i\}$ by making selections from a uniformly distributed set randomly biased according to the *density* function. In any given situation, speed and convenience will usually select one method over the other. The inversion method is 100% efficient in its use of random numbers, but it requires calculating and inverting the distribution function, a task which is sometimes quite difficult. The rejection method does not require a knowledge of the distribution function nor even the absolutely normalized density function, but it does require us to know a reasonable upper bound on the density function; moreover, if the shape of the curve P(x)-versus-x is such that the area of the smallest box enclosing this curve is very much larger than the area under this curve, then the rejection method will be very inefficient.

One other method for generating a set of random numbers will be described in Sec. 2-8, after we have examined the problem of generating random points in more than one dimension.

2.4. Specifying Sets of Random Points

We shall now see how the foregoing ideas concerning the specification and construction of sets of random points in one dimension can be generalized to any number of dimensions. For concreteness, and with no real loss of generality, we shall confine our discussion mainly to the three-dimensional case; here we denote a general point by $\vec{x}=(x,y,z)$ where x, y and z are ordinary real variables. When we speak of a set of random points $\{\vec{x}_i\}\equiv\{(x_i,y_i,z_i)\}$ distributed according to the probability density function $P(\vec{x})\equiv P(x,y,z)$, we mean an inexhaustible set of triplets of real numbers from which we may "draw" sequential elements (x_1,y_1,z_1), (x_2,y_2,z_2), ..., such that

$P(\mathbf{x})d\mathbf{x} \equiv P(x,y,z)dxdydz$

$\equiv$ probability that x_i will lie between x and $x+dx$, and y_i will lie between y and $y+dy$, and z_i will lie between z and $z+dz$. (2.13)

A set of random points $\{(x_i,y_i,z_i)\}$ is completely characterized by its density function $P(x,y,z)$. However, it is often convenient to introduce a number of "lesser" density functions which characterize only certain particular features of the distribution. For example, we may define the *contracted density functions* $P(x,y)$ and $P(x)$ by

$P(x,y)dxdy \equiv$ probability that x_i will lie between x and $x+dx$, and y_i will lie between y and $y+dy$, regardless of where z_i lies. (2.14a)

and

$P(x)dx \equiv$ probability that x_i will lie between x and $x+dx$, regardless of where y_i and z_i lie. (2.14b)

In a similar way we may also define the contracted density functions $P(y,z)$, $P(x,z)$, $P(y)$ and $P(z)$. Of course, the *functional forms* of these contracted density functions will in general all be different; e.g., $P(y,z)$ is generally not the same function of y and z as $P(x,y)$ is of x and y, and $P(z)$ is generally not the same function of z as $P(x)$ is of x. Nevertheless, we shall avoid a cumbersome subscripting of these P-functions, and trust that our meaning will always be clear from context.

It is easy to obtain expressions for the contracted density functions in terms of the full density function P(x,y,z), simply by invoking the addition theorem for probabilities. Thus, the probability in (2.14a) is obtained simply by summing (integrating) the probability in (2.13) over all dz-intervals, and the probability in (2.14b) is obtained by further summing over all dy-intervals:

$$P(x,y) = \int_{-\infty}^{\infty} dz' P(x,y,z') \tag{2.15a}$$

$$P(x) = \int_{-\infty}^{\infty} dy' \int_{-\infty}^{\infty} dz' P(x,y',z') \tag{2.15b}$$

Of course, if we sum (2.13) over <u>all</u> xyz-space, we should get unity (certainty), just as in (2.3):

$$\int_{-\infty}^{\infty} dx' \int_{-\infty}^{\infty} dy' \int_{-\infty}^{\infty} dz' P(x',y',z') = 1 \tag{2.16}$$

In addition to the contracted density functions defined in (2.14), we will also make use of various <u>conditional density functions</u>, which are defined as follows:

$P(y,z|x)dydz \equiv$ probability that y_i will lie between y and y+dy, and z_i will lie between z and z+dz, given that x_i=x. (2.17a)

$P(y|x)dy \equiv$ probability that y_i will lie between y and y+dy, given that x_i=x, regardless of where z_i lies. (2.17b)

$$P(z|x,y)dz \equiv \text{probability that } z_i \text{ will lie between } z \text{ and } z+dz, \text{ given that } x_i=x \text{ and } y_i=y. \tag{2.17c}$$

We read $P(y,z|x)$ as "P of y and z conditioned on x", $P(y|x)$ as "P of y conditioned on x", and $P(z|x,y)$ as "P of z conditioned on x and y". We may obviously introduce six more conditional density functions with different arrangements of the variables with respect to the vertical slash — e.g., $P(x,z|y)$, $P(x|y)$, etc. However, it should be clearly understood that all these conditional density functions are generally different functional forms—e.g., $P(x|y)$ is not the same function of x and y as $P(x|z)$ is of x and z, etc.

As with the contracted density functions in (2.14), the conditional density functions in (2.17) are completely determined by the form of the full density function $P(x,y,z)$. We may derive the expressions for the conditional density functions in (2.17) as follows: Applying the multiplication theorem for probabilities to the probabilities defined in (2.14b) and (2.17a), we see that

$$P(x)dx \cdot P(y,z|x)dydz = P(x,y,z)dxdydz$$

Therefore,

$$P(y,z|x) = P(x,y,z)/P(x)$$

or, with (2.15b),

$$P(y,z|x) = P(x,y,z)\Big/\int_{-\infty}^{\infty}dy'\int_{-\infty}^{\infty}dz'P(x,y',z') \tag{2.18a}$$

Now treating x as a fixed parameter, the addition theorem for probabilities yields the following relation between the probabilities defined in (2.17a) and (2.17b):

$$P(y|x)dy = \int_{-\infty}^{\infty} dz' dy P(y,z'|x)$$

Inserting (2.18a) yields for $P(y|x)$ the formula

$$P(y|x) = \int_{-\infty}^{\infty} dz' P(x,y,z') \Big/ \int_{-\infty}^{\infty} dy' \int_{-\infty}^{\infty} dz' P(x,y',z') \qquad (2.18b)$$

Finally, again treating x as fixed and applying the multiplication theorem to the probabilities defined in (2.17b) and (2.17c), we see that

$$P(y|x)dy \cdot P(z|x,y)dz = P(y,z|x)dydz$$

Therefore,

$$P(z|x,y) = P(y,z|x)/P(y|x)$$

or, inserting (2.18a) and (2.18b),

$$P(z|x,y) = P(x,y,z) \Big/ \int_{-\infty}^{\infty} dz' P(x,y,z') \qquad (2.18c)$$

It will be observed from the explicit formulae for the two-dimensional density functions $P(x,y)$ in (2.15a) and $P(y,z|x)$ in (2.18a) that each is correctly normalized:

$$\int_{-\infty}^{\infty} dx' \int_{-\infty}^{\infty} dy' P(x',y') = \int_{-\infty}^{\infty} dy' \int_{-\infty}^{\infty} dz' P(y',z'|x) = 1 \qquad (2.19)$$

Similarly, it will be observed from the explicit formulae for the one-dimensional density functions $P(x)$ in (2.15b), $P(y|x)$ in (2.18b) and $P(z|x,y)$ in (2.18c) that they are also correctly normalized:

$$\int_{-\infty}^{\infty}dx'P(x') = \int_{-\infty}^{\infty}dy'P(y'|x) = \int_{-\infty}^{\infty}dz'P(z'|x,y) = 1 \qquad (2.20)$$

The formulae for these one-dimensional density functions will also be observed to imply the following important relation [cf. (2.15b), (2.18b) and (2.18c)]:

$$P(x,y,z) = P(x)P(y|x)P(z|x,y) \qquad (2.21a)$$

The physical meaning of this equation is best seen by writing it in the form

$$P(x,y,z)dxdydz = P(x)dx\cdot P(y|x)dy\cdot P(z|x,y)dz \qquad (2.21b)$$

which says that the probability for simultaneously finding x_i, y_i and z_i in the respective intervals $(x,x+dx)$, $(y,y+dy)$ and $(z,z+dz)$ is equal to the <u>product</u> of: (<u>i</u>) the probability for finding x_i in $(x,x+dx)$, times (<u>ii</u>) the probability for finding y_i in $(y,y+dy)$ given that x_i=x, times (<u>iii</u>) the probability for finding z_i in $(z,z+dz)$ given that x_i=x and y_i=y. In other words, (2.21b) is really a consequence of the multiplication theorem for probabilities. We shall refer to the act of expressing the full three-variable density function P(x,y,z) as a product of three one-variable density functions as "conditioning P(x,y,z)". The fact that we have derived explicit formulae for the three one-dimensional density functions in (2.21), namely

$$P(x) = \int_{-\infty}^{\infty}dy'\int_{-\infty}^{\infty}dz'P(x,y',z') \qquad (2.22a)$$

$$P(y|x) = \int_{-\infty}^{\infty}dz'P(x,y,z') \Big/ \int_{-\infty}^{\infty}dy'\int_{-\infty}^{\infty}dz'P(x,y',z') \qquad (2.22b)$$

$$P(z|x,y) = P(x,y,z)\Big/\int_{-\infty}^{\infty}dz'P(x,y,z') \qquad (2.22c)$$

proves that in principle it is always possible to condition $P(x,y,z)$. Indeed, it is always possible to carry out the conditioning with respect to <u>any ordering</u> of the variables; e.g., we could condition $P(x,y,z)$ as $P(y)P(x|y)P(z|x,y)$ or as $P(y)P(z|y)P(x|y,z)$, etc.

Since $P(x)$, $P(y|x)$ and $P(z|x,y)$ in (2.22) are <u>one</u>-dimensional density functions, then we can introduce in analogy with (2.2) their associated distribution functions $F(x)$, $F(y|x)$ and $F(z|x,y)$:

$$F(x) \equiv \int_{-\infty}^{x}P(x')dx' \qquad (2.23a)$$

$$F(y|x) \equiv \int_{-\infty}^{y}P(y'|x)dy' \qquad (2.23b)$$

$$F(z|x,y) \equiv \int_{-\infty}^{z}P(z'|x,y)dz' \qquad (2.23c)$$

Thus, for example, $F(y|x)$ is the probability that y_i will be less than y, given that $x_i=x$, regardless of the value of z_i.

We may also define a three-variable distribution function $F(x,y,z)$ by

$$F(x,y,z) \equiv \int_{-\infty}^{x}dx'\int_{-\infty}^{y}dy'\int_{-\infty}^{z}dz'P(x',y',z') \qquad (2.24)$$

Evidently, $F(x,y,z)$ is the probability that a randomly selected element (x_i,y_i,z_i) will simultaneously have $x_i \leq x$, $y_i \leq y$ and $z_i \leq z$. In Monte Carlo applications, though, distribution functions with more than one argument are not of much use, for we recall that $F(x)$ in (2.2) is chiefly of interest because of the role which F^{-1} plays in the inversion generating method. However, $F(x,y,z)$ in (2.24) is evidently a mapping from 3-space into 1-space, and hence does not have an inverse. The one-variable distribution functions in (2.23), on the other hand, do nave unique inverses, and they will play an important role in the generalization of the inversion generating method, as will be seen in the next section.

Suppose a given set of random points $\{(x_i,y_i,z_i)\}$ with density function $P_1(x,y,z)$ is transformed into a new set of random points $\{(u_i,v_i,w_i)\}$ by the transformation

$$\left.\begin{aligned} u &= U(x,y,z) \\ v &= V(x,y,z) \\ w &= W(x,y,z) \end{aligned}\right\} \qquad (\ 25)$$

What will be the density function $P_2(u,v,w)$ of the set $\{(u_i,v_i,w_i)\}$? If dudvdw is the image of the volume element dxdydz under the transformation (2.25), then clearly the probability for finding a point (u_i,v_i,w_i) inside dudvdw is the same as the probability for finding a point (x_i,y_i,z_i) inside dxdydz. Hence, in analogy with (2.6), we have

$$P_2(u,v,w)dudvdw = P_1(x,y,z)dxdydz \qquad (2.26)$$

The mathematical statement of the fact that the volume element dudvdw centered at (u,v,w) is the image of the volume element dxdydz at (x,y,z) is simpiy Eq. (2.25) together with [cf. (2.5b)]

$$dudvdw = \left|\frac{\partial(u,v,w)}{\partial(x,y,z)}\right| dxdydz \qquad (2.27a)$$

Here,

$$\frac{\partial(u,v,w)}{\partial(x,y,z)} \equiv \begin{vmatrix} \frac{\partial u}{\partial x} & \frac{\partial v}{\partial x} & \frac{\partial w}{\partial x} \\ \frac{\partial u}{\partial y} & \frac{\partial v}{\partial y} & \frac{\partial w}{\partial y} \\ \frac{\partial u}{\partial z} & \frac{\partial v}{\partial z} & \frac{\partial w}{\partial z} \end{vmatrix} \tag{2.27b}$$

is the Jacobian of the transformation (2.25), in which it is understood that the partial derivatives are all evaluated via (2.25) at the point (x,y,z) under consideration. [Readers who are not altogether familiar with Jacobians and their significance may find the short, heuristic discussion given in Appendix B helpful.] With (2.27a), (2.26) implies that the density function of the transformed set $\{(u_i,v_i,w_i)\}$ is

$$P_2(u,v,w) = P_1(x,y,z) \Bigg/ \left|\frac{\partial(u,v,w)}{\partial(x,y,z)}\right| \tag{2.28a}$$

or equivalently [cf. (B. 8)]

$$P_2(u,v,w) = P_1(x,y,z) \cdot \left|\frac{\partial(x,y,z)}{\partial(u,v,w)}\right| \tag{2.28b}$$

where x,y and z are <u>now</u> to be regarded as functions of u,v and w through the <u>inverse</u> of (2.25). If the transformation (2.25) is not strictly one-to-one, so that a given dudvdw element is populated by several dxdydz elements, then the right sides of (2.26) and (2.28) will have to be summed over all contributing dxdydz elements.

2-5. The Generalized Inversion Method

Let us now see how the inversion method for generating random numbers x_i according to a given density function P(x) can be generalized to generate random triplets (x_i, y_i, z_i) according to a given density function $P(x,y,z)$. As already mentioned, the noninvertibility of the distribution function $F(x,y,z)$ precludes its use in a formula of the type (2.9). Instead, we proceed as follows:

> Generalized Inversion Method: First, condition the given density function $P(x,y,z)$ in the form $P(x)P(y|x)P(z|x,y)$ [cf. (2.21) and (2.22)], and calculate the corresponding one-dimensional distribution functions $F(x)$, $F(y|x)$ and $F(z|x,y)$ [cf. (2.23)]. Then, with r_{1i}, r_{2i} and r_{3i} three independent random numbers drawn from the set $\{r_i\}$, first obtain x_i by solving (inverting)
>
> $$r_{1i} = F(x_i) \tag{2.29a}$$
>
> then obtain y_i by solving (inverting)
>
> $$r_{2i} = F(y_i|x_i) \tag{2.29b}$$
>
> where x_i is the value found in (2.29a), and finally, obtain z_i by solving (inverting)
>
> $$r_{3i} = F(z_i|x_i, y_i) \tag{2.29c}$$
>
> where x_i and y_i are the values found in (2.29a) and (2.29b), respectively.

That the set $\{(x_i, y_i, z_i)\}$ constructed according to the foregoing procedure actually has $P(x,y,z)$ as its density function can be proved as follows: First note that, in picking three random numbers r_{1i}, r_{2i} and r_{3i} from the set of random numbers $\{r_i\}$ distributed uniformly in the unit interval, we are essentially picking one random point (r_{1i},r_{2i},r_{3i}) from the set of random points distributed uniformly over the unit cube in $r_1r_2r_3$-space. That is, since the probability for simultaneously finding r_{1i} in (r_1,r_1+dr_1), r_{2i} in (r_2,r_2+dr_2), and r_{3i} in (r_3,r_3+dr_3) is just

$$P(r_1)dr_1 \cdot P(r_2)dr_2 \cdot P(r_3)dr_3$$

where P(r) is given by (2.8a), then the probability density function $P_1(r_1,r_2\ r_3)$ for the set of random triplets $\{(r_{1i},r_{2i},r_{3i})\}$ is

$$P_1(r_1,r_2,r_3) = P(r_1)P(r_2)P(r_3) = \begin{cases} 1, & \text{if } 0 \leqslant r_j \leqslant 1,\ j=1,2,3 \\ 0, & \text{otherwise} \end{cases}$$

Now, regarding (2.29), or the inverse thereof, as a transformation which carries each (r_{1i},r_{2i},r_{3i})-point into a (x_i,y_i,z_i)-point, it follows from (2.28b) that the density function $P_2(x,y,z)$ of the set $\{(x_i,y_i,z_i)\}$ is

$$P_2(x,y,z) = P_1(r_1,r_2,r_3) \cdot \left| \frac{\partial(r_1,r_2,r_3)}{\partial(x,y,z)} \right|$$

$$= 1 \cdot \left\| \begin{matrix} \frac{\partial r_1}{\partial x} & \frac{\partial r_2}{\partial x} & \frac{\partial r_3}{\partial x} \\ \frac{\partial r_1}{\partial y} & \frac{\partial r_2}{\partial y} & \frac{\partial r_3}{\partial y} \\ \frac{\partial r_1}{\partial z} & \frac{\partial r_2}{\partial z} & \frac{\partial r_3}{\partial z} \end{matrix} \right\|$$

Now, from (2.29a), r_1 is independent of y and z, and from (2.29b), r_2 is independent of z. Hence, all elements of the Jacobian determinant below the main diagonal vanish, so

$$P_2(x,y,z) = \left|\frac{\partial r_1}{\partial x}\cdot\frac{\partial r_2}{\partial y}\cdot\frac{\partial r_3}{\partial z}\right|$$

$$= \left|\frac{\partial}{\partial x}F(x)\cdot\frac{\partial}{\partial y}F(y|x)\cdot\frac{\partial}{\partial z}F(z|x,y)\right| \qquad \text{[by (2.29)]}$$

$$= \left|P(x)\cdot P(y|x)\cdot P(z|x,y)\right| \qquad \text{[by (2.23)]}$$

$$= P(x,y,z) \qquad \text{[by (2.21)]}$$

which establishes the desired result.

From the point of view of the foregoing proof, the generalized inversion formulae in (2.29) produce the desired results because these formulae constitute a transformation from $r_1r_2r_3$-space to xyz-space which has the rather unique property that

$$\frac{\partial(r_1,r_2,r_3)}{\partial(x,y,z)} = P(x,y,z) \qquad (2.30)$$

From a less formal point of view, however, it is clear that the three-dimensional inversion method i_ nothing more than three successive applications of the one-dimensional inversion method to the conditioned form of the density function. That is, (2.29a) generates a random number x_i according to P(x), (2.29b) generates a random number y_i according to $P(y|x_i)$, and (2.29c) generates a random number z_i according to $P(z|x_i,y_i)$.

Thus, once one appreciates the significance of the conditioned form of the density function in (2.21), the generalized inversion method presented above is intuitively quite plausible.

It should be noted that one has considerable flexibility in applying the generalized inversion method. Thus, if one or more of the distribution functions $F(x)$, $F(y|x)$ and $F(z|x,y)$ are intractible, one can try to condition $P(x,y,z)$ in another form, say, $P(y)\cdot P(z|y)\cdot P(x|y,z)$, and thus work with the different distribution functions $F(y)$, $F(z|y)$ and $F(x|y,z)$. Alternatively, it will be observed that any or all of the three successive steps in (2.29a), (2.29b) and (2.29c) could actually be carried out by applying the one-dimensional rejection method. For example, once x_i has been picked according to (2.29a), one could replace (2.29b) by an application of the one-dimensional rejection method to generate a random point y_i according to the density function $P(y|x_i)$, and then proceed as usual with (2.29c). However, we shall regard such applications of the one-dimensional rejection method as still falling under the scope of the "generalized inversion method", and reserve the term "generalized rejection method" for a procedure to be described later.

2-6. Generating Uniformly Distributed Random Points

A very important application of the generalized inversion method is the generating of random points from a uniform distribution inside some given region Ω. Suppose for now that Ω can be specified in the following way:[†]

$$\Omega = \{(x,y,z) \| a_1 \leqslant x \leqslant b_1,\ a_2(x) \leqslant y \leqslant b_2(x),\ a_3(x,y) \leqslant z \leqslant b_3(x,y)\} \quad (2.31a)$$

The volume $|\Omega|$ of the region Ω is thus given by the integral

$$|\Omega| = \int_{a_1}^{b_1} dx' \int_{a_2(x')}^{b_2(x')} dy'[b_3(x',y') - a_3(x',y')] \qquad (2.31b)$$

where the integrand of course represents the result of the trivial z'-integration.

The density function defining a uniform distribution of points inside Ω is

$$P(x,y,z) = \begin{cases} 1/|\Omega|, & \text{for } (x,y,z)\varepsilon\Omega \\ 0 \quad , & \text{for } (x,y,z)\not\varepsilon\Omega \end{cases} \qquad (2.32)$$

To apply the generalized inversion method to generate random points according to this P(x,y,z), we must evidently condition P(x,y,z) in the manner of (2.21). This is not in general a trivial task, because of the boundaries of Ω. Thus, inserting (2.32) into (2.22), we find for the one-variable density functions

$$P(x) = |\Omega|^{-1} \int_{a_2(x)}^{b_2(x)} dy'[b_3(x,y') - a_3(x,y')], \quad a_1 \leqslant x \leqslant b_1 \qquad (2.33a)$$

[†] Eq. (2.31a) is to be read "Ω is the set of all points (x,y,z) for which $a_1 \leqslant x \leqslant b_1$, $a_2(x) \leqslant y \leqslant b_2(x)$ and $a_3(x,y) \leqslant z \leqslant b_3(x,y)$".

$$P(y|x) = [b_3(x,y) - a_3(x,y)] \Big/ \int_{a_2(x)}^{b_2(x)} dy'[b_3(x,y') - a_3(x,y')],$$

$$a_2(x) \leqslant y \leqslant b_2(x) \qquad (2.33b)$$

$$P(z|x,y) = 1 \big/ [b_3(x,y) - a_3(x,y)], \quad a_3(x,y) \leqslant z \leqslant b_3(x,y) \qquad (2.33c)$$

The corresponding one-variable distribution functions in (2.23) are therefore given by

$$F(x) = |\Omega|^{-1} \int_{a_1}^{x} dx' \int_{a_2(x')}^{b_2(x')} dy'[b_3(x',y') - a_3(x',y')], \quad a_1 \leqslant x \leqslant b_1 \qquad (2.34a)$$

$$F(y|x) = \int_{a_2(x)}^{y} dy'[b_3(x,y') - a_3(x,y')] \Big/ \int_{a_2(x)}^{b_2(x)} dy'$$

$$[b_3(x,y') - a_3(x,y')], \quad a_2(x) \leqslant y \leqslant b_2(x) \qquad (2.34b)$$

$$F(z|x,y) = [z - a_3(x,y)] \big/ [b_3(x,y) - a_3(x,y)],$$

$$a_3(x,y) \leqslant z \leqslant b_3(x,y) \qquad (2.34c)$$

Thus, to generate a random point uniformly inside Ω by the generalized inversion method, we must insert the above distribution functions into (2.29), and solve successively for x_i, y_i and z_i. Depending upon the shape of Ω --i.e., depending upon the boundary functions $a_2(x)$, $b_2(x)$, $a_3(x,y)$, $b_3(x,y)$--this may be a very easy task or a very difficult task.

The easiest case is realized when Ω is a "box", with the boundary functions a_i and b_i all constants. In this case (2.31b) gives

$$|\Omega| = (b_1-a_1)(b_2-a_2)(b_3-a_3)$$

The one-variable density functions in (2.33) become

$$P(x) = 1/(b_1-a_1), \quad a_1 \leqslant x \leqslant b_1$$

$$P(y|x) = 1/(b_2-a_2), \quad a_2 \leqslant y \leqslant b_2$$

$$P(z|x,y) = 1/(b_3-a_3), \quad a_3 \leqslant z \leqslant b_3$$

and the corresponding one-variable distribution functions in (2.34) become

$$F(x) = (x-a_1)/(b_1-a_1), \quad a_1 \leqslant x \leqslant b_1$$

$$F(y|x) = (y-a_2)/(b_2-a_2), \quad a_2 \leqslant y \leqslant b_2$$

$$F(z|x,y) = (z-a_3)/(b_3-a_3), \quad a_3 \leqslant z \leqslant b_3$$

Inserting these distribution functions into (2.29) and inverting, we obtain the following algorithm for generating a random point (x_i,y_i,z_i) from a uniform distribution inside the box $\{a_1 \leqslant x \leqslant b_1, a_2 \leqslant y \leqslant b_2, a_3 \leqslant z \leqslant b_3\}$:

$$x_i = a_1 + (b_1-a_1)r_{1i} \tag{2.35a}$$

$$y_i = a_2 + (b_2-a_2)r_{2i} \tag{2.35b}$$

$$z_i = a_3 + (b_3-a_3)r_{3i} \tag{2.35c}$$

Here, r_{1i}, r_{2i} and r_{3i} are independent random numbers from a uniform distribution in the unit interval. Eqs. (2.35) are precisely what we should expect on the basis of the rule in (2.10c): we simply generate each coordinate independently from a uniform distribution along the corresponding edge of the box.

When the boundary functions a_i and b_i are <u>not</u> all constants, so that Ω is not a box, then the one-dimensional distribution functions must be calculated according to (2.34), inserted into (2.29), and inverted. It is important to realize that, in this general case, one will <u>not</u> obtain equations having the simple form (2.35). That is, although (2.34c) and (2.29c) will indeed produce an equation like (2.35c) with a_3 and b_3 replaced by $a_3(x_i,y_i)$ and $b_3(x_i,y_i)$, (2.34b) and (2.29b) will not produce an equation like (2.35b) with a_2 and b_2 replaced by $a_2(x_i)$ and $b_2(x_i)$, and (2.34a) and (2.29a) will not produce (2.35a). To put it differently, although $P(z|x,y)$ in (2.33c) indeed describes, for fixed x and y, a uniform distribution in z, $P(y|x)$ in (2.33b) does not describe, for fixed x, a uniform distribution in y, and $P(x)$ in (2.33a) does not describe a uniform distribution in x. The point here is that the correct version of the algorithm in (2.35) for non-box regions Ω cannot be easily intuited <u>a priori</u>.

It is in principle always possible to apply the generalized inversion method to generate random points uniformly inside a given region Ω, provided Ω is defined by means of boundary functions a_i and b_i as in (2.31a). In practice, though, the calculation of the one-variable distribution functions in (2.34) and their subsequent inversion often prove to be prohibitively difficult. Furthermore, it often happens that the volume Ω is not defined through boundary functions of the kinds in (2.31a), but rather through one or more inequalities between various functions of the coordinates. In such situations, it is sometimes feasible to proceed in the following alternate way: Choose a box-like region Σ which completely encloses the given region Ω. Generate random points uniformly inside Σ according to the procedure described in connection with (2.35), but keep <u>only</u> those points which happen to <u>also</u> lie inside Ω. Clearly, the subset of "kept" points will be distributed randomly and uniformly inside Ω. This simple procedure has the advantage that one can apply it without having to calculate and invert the various one-dimensional distribution functions. Furthermore, one does not even need to know the boundary functions a_i and b_i in (2.31a); one only needs to be able to decide whether or not a given point in Σ lies inside Ω. The only possible drawback to this method is its efficiency. Clearly, the approximate fraction of uniformly distributed random points inside Σ which also lie inside Ω will be the ratio of the volumes, $|\Omega|/|\Sigma|$. If this ratio is very small--i.e., if Ω is so shaped that its volume is much smaller than the smallest box Σ which can be fitted around Ω--then this method for generating random points uniformly inside Ω will be correspondingly inefficient.

We shall now illustrate the foregoing two procedures for generating random points uniformly inside non-box regions by considering the following two-dimensional problem: Let Ω be the region in the xy-plane which is bounded by the x-axis, the line $x=1$, and the curve $y=x^n$, where n is a fixed, positive integer. A sketch of Ω is shown in Fig. 4.

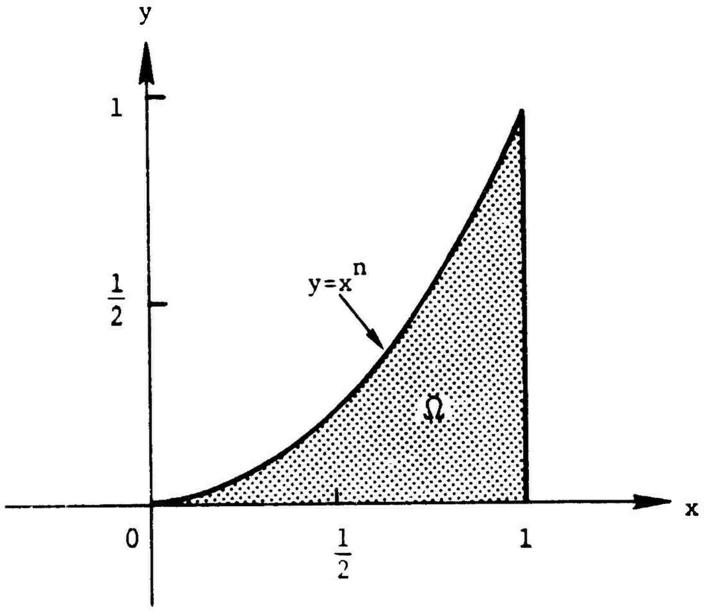

FIGURE 4. An example of a two-dimensional region Ω.

Suppose we wish to generate random points (x_i,y_i) uniformly inside Ω.

One way of proceeding would be to generate random points uniformly inside the unit square Σ, and then keep only those points which happen to fall inside Ω. Letting r_{1i} and r_{2i} be two independent random numbers from a unifor. iistribution in the unit interval, the generating algorithm is evidently

$$\left.\begin{array}{l} x_i = r_{1i} \\ y_i = r_{2i} \\ \text{keep only if } y_i \leqslant x_i^n \end{array}\right\} \qquad (2.36)$$

Since the volume of the unit square is $|\Sigma|=1$ and the volume of Ω is

$$|\Omega| = \int_o^1 dx \int_o^{x^n} dy = \int_o^1 x^n dx = \frac{1}{n+1} \qquad (2.37)$$

then the efficiency of this method is

$$|\Omega|/|\Sigma| = \frac{1}{n+1} \qquad (2.38)$$

For small values of n this method would nct be too bad; e.g., for n=1, Ω would be a simple triangle, and half the points generated inside Σ would be kept. However, if n is very large this method would evidently not be satisfactory. Let us see how we could generate the points inside Ω directly using the generalized inversion method.

We wish to generate random points (x_i,y_i) according to the density function

$$P(x,y) = \begin{cases} 1/|\Omega|, & \text{for } (x,y)\varepsilon\Omega \\ 0 \;\; , & \text{for } (x,y)\notin\Omega \end{cases} \qquad (2.39)$$

where Ω is the volume shown in Fig. 4, and $|\Omega|$ is given by (2.37). We first "condition" this density function in the form

$$P(x,y) = P(x)P(y|x) \qquad (2.40)$$

The one-variable density functions P(x) and P(y|x) are given by [cf. (2.22)]

$$P(x) \equiv \int_0^{x^n} P(x,y')dy' = (n+1)x^n, \quad 0\leqslant x\leqslant 1 \qquad (2.41a)$$

$$P(y|x) \equiv \frac{P(x,y)}{\int_0^{x^n} P(x,y')dy'} = x^{-n}, \quad 0\leqslant y\leqslant x^n \qquad (2.41b)$$

and the corresponding one-variable distribution functions are given by [cf. (2.23)]

$$F(x) \equiv \int_0^x P(x')dx' = x^{n+1}, \quad 0\leqslant x\leqslant 1 \qquad (2.42a)$$

$$F(y|x) \equiv \int_0^y P(y'|x)dy' = x^{-n}y, \quad 0\leqslant y\leqslant x' \qquad (2.42b)$$

Then, with r_{1i} and r_{2i} two independent random numbers from a uniform distribution in the unit interval, we put in accordance with (2.29), $r_{1i}=F(x_i)$ and $r_{2i}=F(y_i|x_i)$, and solve successively for x_i and y_i. The result is easily found to be

$$\left.\begin{aligned} x_i &= r_{1i}^{1/(n+1)} \\ y_i &= x_i^n r_{2i} \end{aligned}\right\} \qquad (2.43)$$

Thus, (2.43) is the algorithm whereby one directly generates random points uniformly inside the region Ω, shown in Fig. 4, for any fixed value of n.

It is tempting to try to "improve" the algorithm (2.36) by modifying its second equation to read

$$y_i = x_i^n r_{2i}$$

This would amount to first generating a random coordinate x_i uniformly between 0 and 1, and then a random coordinate y_i uniformly between 0 and x_i^n (rather than between 0 and 1). Clearly this would a tomatically satisfy the inequality in (2.36), so that ever point (x_i,y_i) generated in this way would lie inside Ω. The trouble with this procedure is that the points (x_i,y_i) generated in this way would not cover Ω uniformly. To see this, we need only observe that this method would produce as many points with $x_i<1/?$ as with $x_i>1/2$, implying that the portion of Ω in Fig. 4 to the left of the line $x=1/2$ would contain just as many points as the portion of Ω to the right of this line--a situation clearly inconsistent with a uniform distribution. The only way of first generating an x_i-value and then generating a y_i-value such that (x_i,y_i) is always a random point from a uniform distribution inside Ω, is to proceed according to the algorithm (2.43).

A more elaborate example of using the generalized inversion method to generate random points uniformly inside a given region will be presented in Section 2-10.

2-7. The Generalized Rejection Method

In the preceding section we showed how one could generate random points uniformly inside a given region Ω. Having this ability, it is possible to generate random points inside Ω according to any prescribed density function $P(x,y,z)$ by a straightforward generalization of the one-dimensional rejection method:

> Generalized Rejection Method: We are given a density function $P(x,y,z)$ which vanishes everywhere outside a specified region Ω, and which is bounded by a number B inside Ω. We require a set of random points $\{(x_i',y_i',z_i')\}$ distributed uniformly over Ω, and also an independent set of random numbers $\{r_i\}$ distributed uniformly over the unit interval. To generate a random point (x_i,y_i,z_i) according to the density function $P(x,y,z)$, draw successive pairs of random points r_i and (x_i',y_i',z_i') until the inequality
>
> $$P(x_i',y_i',z_i')/B \geqslant r_i \tag{2.44}$$
>
> is found to be satisfied, whereupon take $(x_i,y_i,z_i)=(x_i',y_i',z_i')$.

The proof for this method is a straightforward generalization of the proof in one dimension, which is given in Appendix A. As with the one-dimensional case, it should be noted that it is only necessary to know

P(x,y,z) and its upper bound B to within a constant factor, because only their ratio is used. In any case, the efficiency of this method is [cf. (2.12)]

$$E = \frac{\iiint_{\Omega} P(x,y,z)dxdydz}{B\cdot|\Omega|} \tag{2.45}$$

so it is desirable to take B equal to the <u>least</u> upper bound P(max) of P(x,y,z) in Ω.

It may be noted that the alternate technique mentioned in the previous section for generating random points uniformly in a non-box region Ω--namely, by picking from a uniform distribution inside an enclosing box Σ all those points which happen to fall inside Ω--is really an application of the generalized rejection method. Thus, we start with a set of random points $\{(x_i',y_i',z_i')\}$ distributed uniformly inside a box Σ which encloses the given region Ω , and we proceed to construct a set $\{(x_i,y_i,z_i)\}$ distributed according to the density function P(x,y,z) in (2.32). The least upper bound for P(x,y,z) in (2.32) is evidently $B = 1/|\Omega|$, so the ratio on the left of (2.44) will be 1 if $(x_i',y_i',z_i') \varepsilon \Omega$ and 0 if $(x_i',y_i',z_i') \not\varepsilon \Omega$. In the former case the inequality in (2.44) will <u>always</u> be satisfied and the trial point will be kept, while in the latter case the inequality in (2.44) will <u>never</u> be satisfied and the trial point will be rejected. In this case there is never any need to draw a random number r_i: the acceptance of the trial point depends ultimately only on whether it lies inside Ω. The efficiency of this method is calculated from (2.45) by replacing

) by Σ, inserting for $P(x,y,z)$ the function in (2.32) and putting $B=1/|\Omega|$; thus,

$$\mathcal{E} = \frac{1}{(1/|\Omega|)\cdot|\Sigma|} = \frac{|\Omega|}{|\Sigma|}$$

just as we expect.

If the set $\{(x_i',y_i',z_i')\}$ used in the generalized rejection method is distributed over Ω according to a (not necessarily uniform) density function $\tilde{P}(x,y,z)$, then the density function of the set $\{(x_i,y_i,z_i)\}$ constructed in accordance with the selection rule (2.44) would be $C\tilde{P}(x,y,z)P(x,y,z)$, C being the appropriate normalization constant. This follows from a straightforward extension to three dimensions of the arguments presented in Appendix A.

2-8. The Contraction Method

We have discussed two general ways of generating random points according to a prescribed probability density function--namely, the inversion method and the rejection method. We shall now describe one more method, which we shall call the "contraction method", for accomplishing this task. This method is applicable whenever the given density function can be regarded as a contracted density function of some higher dimensional distribution which can be easily handled. In its simplest form, the contraction method can be described as follows:

Contraction Method: It is desired to generate a set of random points $\{x_i\}$ according to a given density function $P(x)$, but it is found that neither the inversion nor rejection method offers an efficient way of doing this. However, it is discovered that there exists a density function P(x,y) for which P(x) is the y-contracted density function:

$$P(x) = \int_{-\infty}^{\infty} P(x,y')dy' \tag{2.46}$$

It further happens that, by using either the generalized inversion method or the generalized rejection method, it is possible to generate random pairs $\{(x_i,y_i)\}$ according to P(x,y) rapidly and efficiently. Then, by generating such a set $\{(x_i,y_i)\}$ and simply <u>ignoring</u> the y-coordinates, we have by (2.46) a set of x-coordinates $\{x_i\}$ which are randomly distributed according to P(x).

We may illustrate the potential usefulness of the contraction generating method by considering the following example. Suppose it is desired to generate a set of random numbers $\{y_i\}$ distributed according to the density function

$$P(y) = \begin{cases} (n+1)[1 - y^{1/n}], & \text{for } 0\leqslant y\leqslant 1 \\ 0, & \text{for } y<0 \text{ and } y>1 \end{cases} \tag{2.47}$$

where n is some fixed, <u>large</u> integer. For the inversion method, we can calculate the distribution function easily enough,

$$F(y) \equiv \int_{0}^{y} P(y')dy' = y[1 + n - ny^{1/n}] \tag{2.48}$$

but we observe that the equation $r_i=F(y_i)$ can be inverted, as required by (2.9), only numerically. The rejection method would entail picking a pair of random numbers y_i and r_i uniformly in the unit interval, and [noting that the least upper bound on $P(y)$ is $B=(n+1)$] taking y_i' to be a member of the desired set $\{y_i\}$ if and only if [cf. (2.11)]

$$P(y_i')/B = 1 - y_i'^{1/n} \geqslant r_i \tag{2.49}$$

However, the efficiency of this method is easily calculated from (2.12) to be $E=1/(n+1)$, which is very low under the given specification that n is large.

We now astutely observe that the density function $P(y)$ given in (2.47) coincides with the x-contracted density function of the quantity $P(x,y)$ defined in (2.39),(2.37) and Fig. 4:[†]

$$P(y) = \int_{-\infty}^{\infty} P(x',y)dx' = \int_{y^{1/n}}^{1} (n+1)dx' = (n+1)[1-y^{1/n}]$$

Now we have already found that the algorithm in (2.43) offers a very efficient way of generating random points $\{(x_i,y_i)\}$ according to $P(x,y)$, even if n is very large. Therefore, by first generating a number x_i according to the first of Eqs. (2.43), and then using this x_i-value to generate a number y_i according to the second of Eqs. (2.43), we will have thereby generated a y_i-value according to the density function in (2.47). Of course, we have had to use two random numbers to do this, but this is still a more efficient

[†] Notice in passing that the functional forms of $P(y)$ in (2.47) and $P(x)$ in (2.41a) are indeed quite different, even though both are contracted from the same two-dimensional density function $P(x,y)$.

method for large n than is offered by either the inversion or rejection methods.

Variations on the contraction method are seen to be virtually limitless. For example, one might find that it is a simple matter to generate a set of random points $\{(x_i,y_i,z_i)\}$ according to a density function $P(x,y,z)$ by applying the generalized inversion method, conditioning $P(x,y,z)$ as $P(x)\cdot P(y|x)\cdot P(z|x,y)$ [cf. (2.29)]. Then, by ignoring the x- and y- coordinates we have available a set of random numbers $\{z_i\}$ distributed according to the density function

$$P(z) = \int_{-\infty}^{\infty}dx'\int_{-\infty}^{\infty}dy'P(x',y',z) \qquad (2.50a)$$

and by ignoring the x-coordinates have available a set of random points $\{(y_i,z_i)\}$ distributed according to the density function

$$P(y,z) = \int_{-\infty}^{\infty}dx'P(x',y,z) \qquad (2.50b)$$

and so on.

We thus have at our disposal a variety of techniques which can be used, in conjunction with a given set of random numbers $\{r_i\}$ distributed uniformly in the unit interval, to construct a set of random points $\{\vec{x}_i\}$ distributed according to any prescribed density function $P(\vec{x})$. In the next chapter we shall see how to use such sets of points to numerically estimate definite integrals. We conclude the present chapter by considering two examples, of interest in both statistics and statistical mechanics, which illustrate some of the ways in which one can utilize the random number generating techniques developed in this chapter.

2-9 An Example: The Weighted Gaussian

Consider the problem of generating a set of random numbers $\{x_i\}$ distributed according to the density function

$$P(x;n,a) = \begin{cases} A(n,a)\ x^n \exp(-ax^2), & x \geqslant 0 \\ 0 \quad , & x < 0 \end{cases} \tag{2.51a}$$

where n is any fixed non-negative integer and a is any fixed positive number. The constant A(n,a) is defined so that P(x;n,a) satisfies the normalization condition (2.3); using standard integral tables one finds

$$A(n,a) \equiv \begin{cases} 2\sqrt{\frac{a}{\pi}} & , \quad n=0 \\ \frac{2}{\sqrt{\pi}}\ \frac{2^{n/2} a^{(n+1)/2}}{1 \cdot 3 \cdot 5 \cdots (n-1)} & , \quad n=2,4,6,\ldots \\ 2a^{(n+1)/2} \Big/ \left(\frac{n-1}{2}\right)! & , \quad n=1,3,5,\ldots \end{cases} \tag{2.51b}$$

For n=0, we have $P(x;0,a)=2\sqrt{a/\pi}\exp(-ax^2)$, $x \geqslant 0$, which is often referred to as the Gaussian curve. [More precisely, the Gaussian curve is usually defined as $\sqrt{a/\pi}\exp(-ax^2)$ on the entire x-axis, so our P(x;0,a) is really just half of the Gaussian curve.] By including the factor x^n, $n>0$, we obtain what we shall term a "weighted Gaussian". It is easy to show that P(x;n,a) assumes its maximum value at the point $x=\sqrt{n/2a}$; furthermore, for $n \geqslant 1$, P(x;n,a) tends to 0 as $x \to 0$, and for all n, P(x;n,a) tends to 0 as $x \to \infty$.

If we wish to generate a set of random numbers $\{x_i\}$ according to P(x;n,a) by the ordinary inversion method, we must first calculate the distribution function

$$F(x;n,a) \equiv \int_0^x P(x';n,a)dx'$$

This calculation is rather lengthy for arbitrary n, and is found to yield

$$F(x;n,a) = \begin{cases} \mathrm{erf}(\sqrt{a}x) & , \quad n=0 \\ \mathrm{erf}(\sqrt{a}x) - \frac{2}{\sqrt{\pi}} \sqrt{a}x \exp(-ax^2) \sum_{\nu=1}^{n/2} \frac{(2ax^2)^{\nu-1}}{1\cdot3\cdot5\cdots(2\nu-1)}, & \quad n=2,4,\ldots \\ 1 - \exp(-ax^2) \sum_{\nu=0}^{(n-1)/2} \frac{(ax^2)^\nu}{\nu!} & , \quad n=1,3,\ldots \end{cases} \quad (2.52)$$

where erf(x) is the so-called "error function",

$$\mathrm{erf}(x) \equiv \frac{2}{\sqrt{\pi}} \int_0^x \exp(-t^2)dt, \quad x \geqslant 0 \qquad (2.53)$$

which is tabulated in many mathematical handbooks. It is clear from (2.52) that the task of inverting F(x;n,a) is in general not a trivial matter. This is particularly true for n=0,2,4,..., since erf(x) can be calculated and inverted only by numerical methods. There is in fact only one case for which F(x,n,a) can be easily handled. This is the case n=1, for the equation

$$r_i = F(x_i;1,a) = 1 - \exp(-ax_i^2)$$

can be easily inverted to obtain

$$x_i = \sqrt{\frac{1}{a} \log\left(\frac{1}{1-r_i}\right)} \qquad (2.54a)$$

as the algorithm whereby one constructs from a set of random numbers

$\{r_i\}$ distributed uniformly in the unit interval, a set of random numbers $\{x_i\}$ distributed according to the density function

$$P(x;1,a) = 2ax\,\exp(-ax^2), \quad x\geq 0 \tag{2.54b}$$

A straightforward application of the rejection method to P(x;n,a) is not very satisfactory because P(x;n,a) is non-zero over an infinite interval. Of course, we might simply put P(x;n,a)=0 for all x larger than some large but finite value x_o, but this procedure is rather arbitrary. Moreover, the efficiency of the rejection method is inversely proportional to the length of the interval $(a,b)=(0,x_o)$ over which the initial uniform set $\{x_i'\}$ is taken [see (2.12)], so the larger we take x_o the more inefficient the rejection method becomes.

We shall now derive two different methods for efficiently generating random numbers $\{x_i\}$ according to the n=0 density function,

$$P(x;0,a) = 2\sqrt{a/\pi}\,\exp(-ax^2), \quad x\geq 0 \tag{2.55}$$

We shall then show how one can easily construct, from a given set of random numbers $\{x_i\}$ distributed according to P(x;0,a), another set of random numbers $\{\rho_i\}$ distributed according to $P(\rho;n,a)$ for any integer n>0.

The first method for generating random numbers $\{x_i\}$ according to P(x;0,a) essentially consists of a combination of the contraction and inversion methods, coupled with a suitable transformation of variables. Consider the auxiliary two-dimensional density function P(x,y), defined by

$$P(x,y) \equiv P(x;0,a)\cdot P(y;0,a)$$

$$= \begin{cases} 4\frac{a}{\pi}\exp(-a[x^2+y^2]), & \text{for } x,y\geqslant 0 \\ 0 & , \text{for } x<0 \text{ or } y<0 \end{cases} \tag{2.56}$$

Clearly, the contracted density functions P(x) and P(y) are

$$P(x) \equiv \int_0^\infty P(x,y')dy' = P(x;0,a) \tag{2.57a}$$

$$P(y) \equiv \int_0^\infty P(x',y)dx' = P(y;0,a) \tag{2.57b}$$

so that if we can generate random pairs $\{(x_i,y_i)\}$ according to P(x,y) then the separate coordinate sets $\{x_i\}$ and $\{y_i\}$ will <u>each</u> be a set of random numbers distributed according to the desired density function. Moreover, since

$$P(x|y) \equiv P(x,y)/P(x) = P(x;0,a) = P(x) \tag{2.58a}$$

and

$$P(y|x) \equiv P(x,y)/P(y) = P(y;0,a) = P(y) \tag{2.58b}$$

then for any random point (x_i,y_i), a knowledge of x_i tells us nothing about the possible values of y_i; in other words, the sets $\{x_i\}$ and $\{y_i\}$ derived from the set $\{(x_i,y_i)\}$ are statistically independent of each other. Now, how can we obtain a set $\{(x_i,y_i)\}$ distributed according to P(x,y) in (2.56)? Consider the transformation of variables $(x,y)\rightarrow(\rho,\theta)$ defined by

$$\left.\begin{aligned} x &= \rho\cos\theta \\ y &= \rho\sin\theta \end{aligned}\right\} \tag{2.59}$$

Since $\partial(x,y)/\partial(\rho,\theta)=\rho$, then a distribution of random pairs $\{(x_i,y_i)\}$ with the density function P(x,y) in (2.56) corresponds, under the transformation (2.59), to a distribution of random pairs $\{(\rho_i,\theta_i)\}$ with density function [cf. (2.25)-(2.28)]

$$\tilde{P}(\rho,\theta) = P(x,y)\left|\frac{\partial(x,y)}{\partial(\rho,\theta)}\right| = 4\frac{a}{\pi}\rho\ \exp(-a\rho^2) \tag{2.60}$$

where $0\leq\rho<\infty$ and $0\leq\theta\leq\pi/2$. Conditioning $\tilde{P}(\rho,\theta)$ in the form $\tilde{P}(\rho)\cdot\tilde{P}(\theta|\rho)$, we find

$$\tilde{P}(\rho) \equiv \int_0^{\pi/2}\tilde{P}(\rho,\theta')d\theta' = 2a\rho\ \exp(-a\rho^2) \equiv P(\rho;1,a) \tag{2.61a}$$

$$\tilde{P}(\theta|\rho) \equiv \tilde{P}(\rho,\theta)/\tilde{P}(\rho) = 1/(\pi/2) \tag{2.61b}$$

Now, we have already seen how to generate random numbers ρ_i according to $P(\rho;1,a)$ [cf. (2.54)]; furthermore, it is trivial to generate random numbers θ_i in $(0,\pi/2)$ according to the density function in (2.61b) [cf. (2.10)]. Hence, it is a simple matter to generate random pairs $\{(\rho_i,\theta_i)\}$ according to $\tilde{P}(\rho,\theta)$. Our algorithm for generating random numbers $\{x_i\}$ according to $P(x;0,a)$ is therefore as follows: Letting r_{1i} and r_{2i} denote two random numbers from a uniform distribution in the unit interval, calculate [cf. (2.54) and (2.10)]

$$\rho_i = \sqrt{\frac{1}{a}\ \log\left(\frac{1}{1-r_{1i}}\right)} \tag{2.62a}$$

$$\theta_i = \left(\frac{\pi}{2}\right)r_{2i} \tag{2.62b}$$

Then calculate, in accordance with (2.59),

$$x_i = \rho_i \cos\theta_i \tag{2.63a}$$

$$y_i = \rho_i \sin\theta_i \tag{2.63b}$$

The random pairs $\{(x_i,y_i)\}$ generated in this way will evidently be distributed according to the density function P(x,y) in (2.56). Therefore, by (2.57), the set $\{x_i\}$ will be distributed according to the density function P(x;0,a) and the set $\{y_i\}$ will be distributed according to the density function P(y;0,a). Moreover, because of (2.58) the sets $\{x_i\}$ and $\{y_i\}$ are <u>statistically</u> <u>independent</u>, so that the numbers x_i and y_i calculated from the <u>same</u> ρ_i and θ_i in (2.63) can be used successively without introducing unwanted correlations. Note that this generating method, which operationally involves nothing more than the formulae in (2.62) and (2.63), is actually 100% efficient, in that the two random numbers r_{1i} and r_{2i} from a uniform distribution in the unit interval actually produce two random numbers distributed according to the desired density function P(x;0,a). [Note also that the quantity $(1-r_{1i})$ in (2.62a) can be replaced by r_{1i}, since both are uniformly distributed random numbers in the unit interval.]

We next consider an alternate method of generating random numbers $\{x_i\}$ according to the density function P(x;0,a). This method consists of first introducing a change of variables x→y which transforms the <u>infinite</u> range $0\leq x<\infty$ into the <u>finite</u> range $0<y\leq 1$, and then applying to the transformed density function the one-dimensional rejection technique. [This method is adapted from Fluendy, Ref. 2, p. 77.] The x→y transformation

used here is

$$y = e^{-\sqrt{a}x} \quad \text{or} \quad x = -\sqrt{\frac{1}{a}}\,\log y \tag{2.64}$$

Under this transformation a set of random numbers $\{x_i\}$ distributed over the interval $0 \leqslant x < \infty$ according to $P(x;0,a)$ corresponds to a set of random numbers $\{y_i\}$ distributed over the interval $0 < y \leqslant 1$ according to the density function [see (2.5)- (2.7)]

$$\tilde{P}(y) = P(x;0,a)\left|\frac{dx}{dy}\right|$$

Using (2.64) and (2.55), we easily find

$$\tilde{P}(y) = \frac{2}{\sqrt{\pi}}\, y^{-1}\exp(-\log^2 y), \qquad 0 < y \leqslant 1 \tag{2.65}$$

It is not difficult to show that $\tilde{P}(y)$ assumes its maximum value at $y=1/\sqrt{e}$, and that this maximum value is

$$B = \frac{2}{\sqrt{\pi}}\, e^{1/4} \tag{2.66}$$

Hence, we can generate a random y_i-value according to $\tilde{P}(y)$ by repetitively drawing pairs of random numbers y_i' and r_i from a uniform distribution in the unit interval until the inequality [cf. (2.11)]

$$\tilde{P}(y_i')/B \geqslant r_i$$

or equivalently

$$\left(\frac{1}{2} + \log y_i'\right)^2 \leqslant \log\left(\frac{1}{r_i}\right) \tag{2.67}$$

is found to be satisfied. We then take $y_i = y_i'$, and put

$$x_i = -\sqrt{\frac{1}{a}}\,\log y_i \qquad (2.68)$$

in accordance with the transformation (2.64). The efficiency of this method—i.e., the fraction of the y_i'-values which lead to acceptable y_i-values, and hence acceptable x_i-values—is found from (2.12) to be

$$\mathcal{E} = \frac{\int_0^\infty \tilde{P}(y)dy}{B\cdot(1-0)} = \frac{1}{B} = \frac{\sqrt{\pi}}{2e^{1/4}} \simeq 0.69 \qquad (2.69)$$

This efficiency is quite satisfactory; it implies that roughly 2 out of every 3 y_i'-values tried will be accepted.

We thus have available <u>two</u> methods of rapidly and efficiently generating random numbers $\{x_i\}$ according to the density function $P(x;0,a)$. We shall now show how one may use random numbers distributed according to $P(x;0,a)$ to construct random numbers distributed according to $P(x;n,a)$ for any integer $n>0$. The method is operationally quite simple: If $x_{1i}, x_{2i}, \ldots, x_{n+1,i}$ are $n+1$ numbers drawn at random from a set $\{x_i\}$ whose density function is $P(x;0,a)$, then

$$\rho_i \equiv \sqrt{x_{1i}^2 + x_{2i}^2 + \ldots + x_{n+1,i}^2} \qquad (2.70)$$

will be a random element from a set $\{\rho_i\}$ whose density function is $P(\rho;n,a)$.

To prove the last statement, consider the $(n+1)$-dimensional density function

$$P(x_1, x_2, \ldots, x_{n+1}) \equiv \prod_{j=1}^{n+1} P(x_j;0,a)$$

$$\equiv \begin{cases} 2^{n+1}\left(\frac{a}{\pi}\right)^{(n+1)/2} \exp(-a[x_1^2 + \ldots x_{n+1}^2]), & \text{if all } x_i \geq 0 \\ 0 \quad , & \text{if any } x_i < 0 \end{cases} \qquad (2.71)$$

We may generate a random point $(x_{1i}, x_{2i}, \ldots, x_{n+1,i})$ according to this density function merely by picking each component independently according to the density function $P(x;0,a)$; this follows because, as may be readily seen from (2.71),

$$\begin{aligned} P(x_1) &= P(x_1;0,a) \\ P(x_2|x_1) &= P(x_2;0,a) \\ P(x_3|x_1,x_2) &= P(x_3;0,a) \\ &\text{etc.} \end{aligned}$$

Consider next the transformation of variables $(x_1, x_2, \ldots, x_{n+1}) \rightarrow (\rho, \alpha_1, \alpha_2, \ldots, \alpha_n)$ which corresponds to a change from the Cartesian respresentation of an (n+1)-dimensional vector to a polar representation. Here, ρ is the "length" of the n+1 dimensional vector, while the α_i's are certain angles or cosines thereof. For example, for n=1 we have $\alpha_1 = \theta$, with

$$\left.\begin{aligned} x_1 &= \rho\cos\theta \\ x_2 &= \rho\sin\theta \end{aligned}\right\}, \quad \rho^2 = x_1^2 + x_2^2, \quad \frac{\partial(x_1,x_2)}{\partial(\rho,\theta)} = \rho$$

and for n=2 we have $\alpha_1 = \cos\theta$ and $\alpha_2 = \phi$, with

$$\left.\begin{aligned} x_1 &= \rho\sin\theta\cos\phi \\ x_2 &= \rho\sin\theta\sin\phi \\ x_3 &= \rho\cos\phi \end{aligned}\right\}, \quad \rho^2 = x_1^2 + x_2^2 + x_3^2, \quad \frac{\partial(x_1,x_2,x_3)}{\partial(\rho,\cos\theta,\phi)} = \rho^2$$

In general, the transformation we consider has the properties that

$$\rho^2 = x_1^2 + x_2^2 + \ldots + x_{n+1}^2 \tag{2.72a}$$

and

$$\frac{\partial(x_1,x_2,\ldots,x_{n+1})}{\partial(\rho,\alpha_1,\ldots,\alpha_n)} = \rho^n \tag{2.72b}$$

From (2.25)-(2.28) it follows that a set of random (n+1)-tuples $\{(x_{1i},x_{2i},\ldots,x_{n+1,i})\}$ distributed according to the density function $P(x_1,x_2,\ldots,x_{n+1})$ corresponds to a set of random (n+1)-tuples $\{(\rho_i,\alpha_{1i},\ldots,\alpha_{ni})\}$ distributed according to the density function

$$\begin{aligned}\tilde{P}(\rho,\alpha_1,\ldots,\alpha_n) &= P(x_1,\ldots,x_{n+1})\left|\frac{\partial(x_1,x_2,\ldots,x_{n+1})}{\partial(\rho,\alpha_1,\ldots,\alpha_n)}\right| \\ &= 2^{n+1}\left(\frac{a}{\pi}\right)^{(n+1)/2} \exp(-a\rho^2)\rho^n \qquad (2.73) \\ &\equiv C(n,a)\rho^n\exp(-a\rho^2)\end{aligned}$$

The fully contracted ρ-density function is therefore

$$\begin{aligned}\tilde{P}(\rho) &= \int d\alpha_1\ldots\int d\alpha_n\tilde{P}(\rho,\alpha_1,\ldots,\alpha_n) \\ &= C'(n,a)\rho^n\exp(-a\rho^2)\end{aligned}$$

or

$$\tilde{P}(\rho) = A(n,a)\rho^n\exp(-a\rho^2) \equiv P(\rho;n,a) \tag{2.74}$$

Here, the second equality follows from the fact that $\tilde{P}(\rho,\alpha_1,\ldots,\alpha_n)$ is independent of each α_n, and the last equality follows by simply recognizing that $\tilde{P}(\rho)$ must in any case be correctly normalized. Hence, we have shown that the quantity ρ defined in (2.72a) is distributed according to $P(\rho;n,a)$. This establishes the simple construction algorithm (2.70).

Actually, the algorithm (2.70) is merely a generalization of a familiar result in statistical mechanics: For gas molecules in thermal equilibrium, each Cartesian component v_i of the molecular velocity $\vec{v}$

is distributed according to the density function $\exp(-mv_i^2/2kT)$; consequently the molecular speed $v=(v_1^2+v_2^2+v_3^2)^{1/2}$ is distributed according to $v^2\exp(-mv^2/2kT)$.

In conclusion, we see that we may generate random numbers $\{x_i\}$ according to the weighted Gaussian function $P(x;n,a)$ in (2.51) either by numerically inverting the distribution function $F(x;n,a)$ in (2.52), or by first generating random numbers according to $P(x;0,a)$ via either (2.62)-(2.63) or (2.67)-(2.68) and then using (2.70).

2-10. An Example: Uniform Distribution of Non-overlapping Rods on a Line

Consider the problem of distributing N line segments or "rods", each of length a, randomly and uniformly inside the x-axis interval (0,L), subject to the constraint that none of the rods overlap. We assume that

$$L > Na \tag{2.75}$$

so that the interval (0,L) is indeed large enough to accommodate all the rods.

One way of proceeding on this problem would be to scatter the rods randomly, uniformly and independently inside the interval (0,L) until we come by chance upon a configuration in which none of the rods overlap. In this approach, we first draw N random numbers $r_1,r_2,\ldots,r_N$ from the uniform distribution in the unit interval, and we tentatively locate the center of rod k at

$$x_k = \frac{a}{2} + \left(L - a\right)r_k, \qquad k=1,2,\ldots,N \tag{2.76a}$$

The resulting configuration $x_1, x_2, \ldots, x_N$ is then accepted if it is found to satisfy the no-overlap condition

$$|x_k - x_j| > a, \quad \text{all } k \neq j \tag{2.76b}$$

If this condition is not satisfied then the configuration is rejected [the entire configuration, not just those x_k's which are found to violate (2.76b)], and we must try again using a different set of random numbers $r_1, r_2, \ldots, r_N$ from the uniform distribution in the unit interval. This procedure is feasible if it turns out that a reasonable fraction of the configurations generated in (2.76a) actually satisfies (2.76b). As we shall prove later [cf. (2.96)], this fraction is in fact given by

$$\frac{\#\text{ acceptable configurations}}{\#\text{ trial configurations}} = \left(\frac{L - Na}{L - a}\right)^N \tag{2.76c}$$

For N=100, a=1 and L=200 this acceptance ratio is $(100/199)^{100} \simeq 1.3 \times 10^{-30}$, which is clearly too small by any standard.

Since the simple rejection generating method just outlined is not generally feasible, let us try to devise an algorithm based on the inversion generating method. First, though, let us restate the problem in a way which shows clearly that we are in fact trying to generate a "point" randomly and uniformly inside a given "region".

Imagine the rods to be laid out on the x-axis in the interval (0,L) in any non-overlapping configuration. Let the rods be numbered from right to left, so that the nearest rod to the left of rod k is always rod k+1, and let x_k locate the center of rod k [see Fig. 5]. Now regard the N variables $x_1, x_2, \ldots, x_N$ as Cartesian coordinates in an N-dimensional hyperspace. Any point in this hyperspace specifies through the values of its coordinates a "configuration" of the rods; however, not every point in this space will satisfy the requirements that the rods be non-overlapping and that the rods be numbered in order from right to left. Let Ω be defined as the set of all points $(x_1, x_2, \ldots, x_N)$ in the N-dimensional configuration space which do satisfy these two requirements; thus, Ω is defined by [see Fig. 5]

$$\Omega \equiv \{(x_1, x_2, \ldots, x_N) \| x_{k+1} + a < x_k < x_{k-1} - a, \quad k=1,\ldots,N\} \quad (2.77)$$

where x_o and x_{N+1} are defined by

$$x_o \equiv L + a/2 \quad (2.78)$$

$$x_{N+1} \equiv -a/2 \quad (2.79)$$

With (2.78) and (2.79) the conditions $x_1 < x_o - a$ and $x_{N+1} + a < x_N$ in (2.77) become respectively

$$x_1 < L - a/2 \quad \text{and} \quad x_N > a/2$$

which conditions evidently insure that rod 1 lies inside the right boundary and rod N lies inside the left boundary [see Fig. 5].

Simply stated, our problem is to generate a point randomly and uniformly inside the N-dimensional region Ω; that is, we wish to generate a random N-tuple $(x_1, \ldots, x_N)$ according to the density function

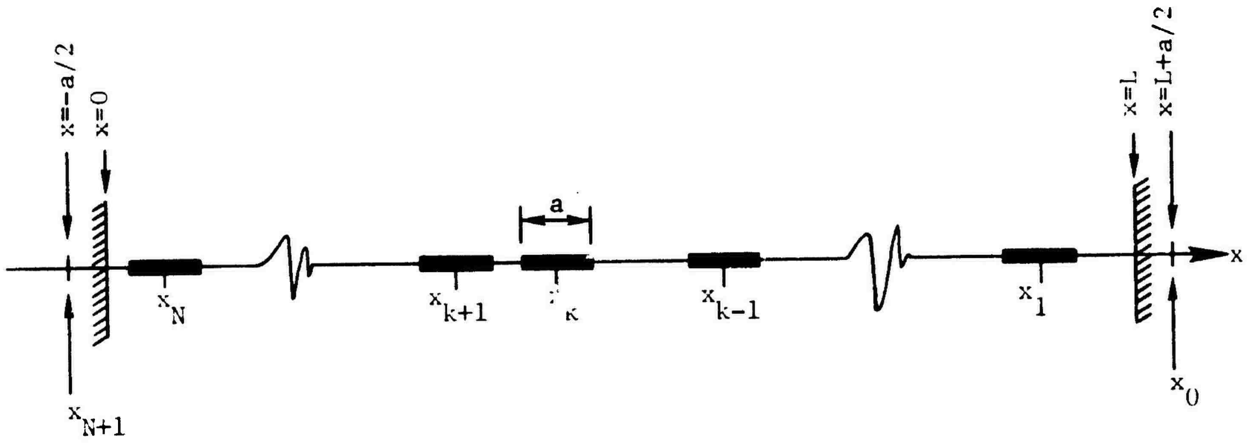

FIGURE 5. N non-overlapping rods of equal length a on a line of length L.

$$P(x_1,\ldots,x_N) = \begin{cases} |\Omega|^{-1}, & \text{if } (x_1,\ldots,x_N)\in\Omega \\ 0\ , & \text{if } (x_1,\ldots,x_N)\notin\Omega \end{cases} \tag{2.80}$$

where $|\Omega|$ is the volume of the region Ω. Our procedure will be to use the generalized inversion method as described in Sections 2-5 and 2-6. For this we shall first need to conduct a detailed analysis of the mathematical properties of the region Ω.

Consider first the variable x_N. From Fig. 5 it is clear that the minimum possible value for x_N is $a/2 \equiv x_{N+1}+a$; the maximum possible value for x_N occurs when the other (N-1) rods are jammed against the right wall, and is $L-(N-1)a-a/2 \equiv L_N$. For any given x_N in the interval $(x_{N+1}+a, L_N)$, the minimum possible value for x_{N-1} is x_N+a; the maximum possible value for x_{N-1} occurs when the remaining (N-2) rods are jammed against the right wall, and is $L-(N-2)a-a/2 \equiv L_{N-1}$. Continuing with this line of reasoning, we see that the volume Ω defined in (2.77) can also be specified in the following way:

$$\Omega = \{(x_1,\ldots,x_N) \,||\, x_{k+1} + a < x_k < L_k, \quad k=1,\ldots,N\} \tag{2.81}$$

where the constants L_k are given by

$$L_k \equiv L - (k-1)a - a/2, \quad k=1,2,\ldots,N+1 \tag{2.82}$$

or equivalently by the recursive formulae

$$L_1 = L - a/2 \tag{2.83a}$$

$$L_{k+1} = L_k - a, \quad k=1,2,\ldots,N \tag{2.83b}$$

[We allow (2.82) and (2.83) to define the quantity L_{N+1} which, although not appearing in (2.81), will be convenient for later formulae.]

The advantage of (2.81) over (2.77) is that it "orders" the coordinaces in the manner of (2.31a), thereby allowing us to employ the techniques outlined in Sec. 2-6.

The volume $|\Omega|$ is given by the N-fold integral

$$|\Omega| = \int_{x_{N+1}+a}^{L_N} dx_N \int_{x_N+a}^{L_{N-1}} dx_{N-1} \ldots \int_{x_2+a}^{L_1} dx_1 \qquad (2.84)$$

The unconditioned density function for x_N is given by [cf. (2.22a), (2.80) and (2.81)]

$$P(x_N) = \int_{x_N+a}^{L_{N-1}} dx_{N-1} \int_{x_{N-1}+a}^{L_{N-2}} dx_{N-2} \cdots \int_{x_2+a}^{L_1} dx_1 |\Omega|^{-1} \qquad (2.85a)$$

The density function for x_k conditioned on x_{k+1}, x_{k+2}, ..., x_N, for $2 \leqslant k \leqslant N-1$ is given by [cf. (2.22b), (2.80), (2.81)]

$$P(x_k|x_{k+1},\ldots,x_N) = \frac{\int_{x_k+a}^{L_{k-1}} dx_{k-1} \ldots \int_{x_2+a}^{L_1} dx_1 |\Omega|^{-1}}{\int_{x_{k+1}+a}^{L_k} dx_k \int_{x_k+a}^{L_{k-1}} dx_{k-1} \ldots \int_{x_2+a}^{L_1} dx_1 |\Omega|^{-1}}, \qquad (2.85b)$$

And the density function for x_1 conditioned on $x_2, x_3, \ldots, x_N$ is given by [cf. (2.22c), (2.80) and (2.81)]

$$P(x_1|x_2,\ldots,x_N) = \frac{|\Omega|^{-1}}{\int_{x_2+a}^{L_1} dx_1 |\Omega|^{-1}} \tag{2.85c}$$

In order to calculate the foregoing quantities, it is convenient to introduce the auxiliary quantities V_0, V_1, ..., V_N defined by

$$V_0 \equiv 1 \tag{2.86a}$$

$$V_k \equiv \int_{x_{k+1}+a}^{L_k} dx_k \int_{x_k+a}^{L_{k-1}} dx_{k-1} \cdots \int_{x_2+a}^{L_1} dx_1, \quad k=1,2,\ldots,N \tag{2.86b}$$

In terms of the quantities V_k we have from (2.84)

$$|\Omega| = V_N \tag{2.87}$$

and from (2.85a) - (2.85c)

$$P(x_k|x_{k+1},\ldots,x_N) = V_{k-1}/V_k, \quad k=1,2,\ldots,N \tag{2.88}$$

provided, for k=N, $P(x_k|x_{k+1},\ldots,x_N)$ is understood to represent $P(x_N)$. Next we shall derive an explicit formula for V_k so that the important quantities above can be calculated.

For k=1 we have

$$V_1 = \int_{x_2+a}^{L_1} dx_1 = L_1 - (x_2+ a) = (L_1- a) - x_2 = L_2- x_2$$

where in the last step we invoked (2.83b). Thus,

$$V_1 = \frac{1}{1!}(L_2- x_2)^1$$

Now suppose that, for any $k \geqslant 1$, V_k is given by

$$V_k = \frac{1}{k!}(L_{k+1} - x_{k+1})^k \tag{2.89}$$

Then from (2.86b) we have

$$\begin{aligned}
V_{k+1} &= \int_{x_{k+2}+a}^{L_{k+1}} dx_{k+1} V_k = \frac{1}{k!}\int_{x_{k+2}+a}^{L_{k+1}} (L_{k+1} - x_{k+1})^k dx_{k+1} \\
&= \frac{1}{k!}\int_{L_{k+1}-(x_{k+2}+a)}^{L_{k+1}-L_{k+1}} z^k(-dz) = \frac{1}{k!}\int_{0}^{L_{k+1}-a-x_{k+2}} z^k dz \\
&= \frac{1}{k!}\,\frac{z^{k+1}}{k+1}\Bigg]_{0}^{L_{k+1}-a-x_{k+2}} = \frac{1}{(k+1)!}\,(L_{k+1} - a - x_{k+2})^{k+1}
\end{aligned}$$

or

$$V_{k+1} = \frac{1}{(k+1)!}(L_{k+2} - x_{k+2})^{k+1}$$

where in the second line we put $z=L_{k+1} - x_{k+1}$ and in the last line we used (2.83b). We have thus proved by induction that (2.89) holds for all $k \geqslant 1$; furthermore, it is seen that for $k=0$ (2.89) gives $V_0=1$, in agreement with the definition in (2.86a). Therefore, (2.89) gives V_k for *all* values of k as defined in (2.86a) and (2.86b).

Inserting (2.89) into (2.87), and invoking the definitions of L_{n+1} and x_{n+1} in (2.82) and (2.79), we find

$$|\Omega| = V_N = (L_{N+1} - x_{N+1})^N/N! = (L - Na - \frac{a}{2} + \frac{a}{2})^N/N!$$

so

$$|\Omega| = (L - Na)^N/N! \tag{2.90}$$

Inserting (2.89) into (2.88) we find

$$P(x_k|x_{k+1},\dots,x_N) = \frac{V_{k-1}}{V_k} = \frac{(L_k - x_k)^{k-1}/(k-1)!}{(L_{k+1} - x_{k+1})^k/k!}$$

or

$$P(x_k|x_{k+1},\dots,x_N) \equiv P(x_k|x_{k+1}) = \frac{k}{(L_{k+1} - x_{k+1})^k}(L_k - x_k)^{k-1} \qquad (2.91)$$

where we have observed that the density function for x_k conditioned on x_{k+1}, x_{k+2}, ..., x_N is in fact independent of x_{k+2}, ..., x_N. The physical reason for this is that the left boundary for rod k is determined solely by the position of rod k+1, and is independent of the positions of all rods to the left of rod k+1. It is of course understood that (2.91) gives $P(x_k|x_{k+1},\dots,x_N)$ only for x_k in the interval $(x_{k+1}+a,\ L_k)$, as prescribed by (2.81); the density function vanishes identically for x_k outside this interval. The formula (2.91) is valid for all k=1,2,...,N provided we keep in mind that, when k=N, the density function is to be regarded as $P(x_N)$.

Our next step is to calculate the one-variable distribution functions $F(x_k|x_{k+1})$ corresponding to the one-variable density functions $P(x_k|x_{k+1})$ in (2.91). Following (2.23) we have

$$F(x_k|x_{k+1}) \equiv \int_{x_{k+1}+a}^{x_k} P(x_k'|x_{k+1})dx_k' = \frac{k}{(L_{k+1}-x_{k+1})^k}\int_{x_{k+1}+a}^{x_k} (L_k - x_k')^{k-1}dx_k'$$

With a change of variable $z=L_k-x_k'$ the integration is easily accomplished. Using (2.83b) the result takes the form

$$F(x_k|x_{k+1}) = 1 - \left(\frac{L_k - x_k}{L_{k+1} - x_{k+1}}\right)^k \tag{2.92}$$

which, as before, holds for x_k in the interval $(x_{k+1}+a, L_k)$ and for all $k=1,2,\ldots,N$. As a check, one can easily verify that $F(x_k|x_{k+1})$ equals zero at the lower limit of x_k [cf. (2.83b)] and equals unity at the upper limit of x_k.

We are now in a position to apply the generalized inversion method described in connection with Eqs. (2.29). Thus, we pick N random numbers $r_1,r_2,\ldots,r_N$ from the uniform distribution in the unit interval, and we solve the equations

$$r_k = F(x_k|x_{k+1}) \tag{2.93}$$

for x_k in the order $k=N,N-1,\ldots,2,1$ as dictated by our conditioning procedure. Substituting (2.92) into (2.93), and recognizing that $1-r_k$ can be replaced by r_k (since both are uniformly distributed random numbers in the unit interval) we have

$$\left(\frac{L_k - x_k}{L_{k+1} - x_{k+1}}\right)^k = r_k$$

Solving for x_k gives the final <u>generating formula</u>:

$$x_k = L_k - (L_{k+1} - x_{k+1})r_k^{1/k} , \quad k=N,N-1,\ldots,1 \tag{2.94}$$

Therefore, the procedure for generating an ordered, non-overlapping but otherwise uniform random configuration of N rods of length a inside the x axis interval (0,L) is as follows: First calculate (and store if more than one configuration is to be generated) the N+2 constants x_{N+1}, L_1, L_2,..., L_{N+1} in accordance with (2.79) and (2.82) or (2.83). Then draw N random numbers r_1, r_2, ..., r_N from the uniform distribution in the unit interval, and compute x_k, the location of the center of rod k, from the formula in (2.94) for the successive k values N,N-1,...,2,1. Notice that (2.94) is to be applied in order of descending k, because the formula for x_k requires a value for x_{k+1}. Essentially, (2.94) just "deals" the rods out on the interval (0,L) from left to right, and our theory assures us that the resultant configuration is acceptable without any further checking.

The rejection technique described in connection with Eqs. (2.76) evidently generates random points uniformly inside the N-dimensional box

$$\Sigma \equiv \{(x_1,\ldots,x_N) \| \frac{a}{2} \leqslant x_k \leqslant L - \frac{a}{2} \text{ , } k=1,\ldots,N\} \qquad (2.95)$$

and then rejects those points which do not also lie inside Ω (which is clearly a subregion of Σ). The efficiency of this method is

$$E = \frac{|\Omega|}{|\Sigma|} = \frac{(L - Na)^N/N!}{(L - a)^N}$$

or

$$E = \frac{1}{N!}\left(\frac{L - Na}{L-a}\right)^N \qquad (2.96)$$

The factor 1/N! in (2.96) simply reflects the fact that the x_k-values generated by this method are generally not ordered according to $x_1 > x_2 > \ldots > x_N$;

hence, the second factor in (2.96) accounts for the no-overlap acceptance ratio, which we have anticipated in (2.76c). If we had a=0, so that we would be generating N points randomly and uniformly inside (0,L), then overlap would clearly not be a problem; however, if it is important to have the points ordered, then the inversion method would still be preferable for large N in that it evidently accomplishes this ordering automatically.

The generating method developed here can be easily extended to generate a random, uniform distribution of equal-length, non-overlapping rods (or more precisely "arcs") around the boundary of a circle. Again denoting the length of each rod by a, it will be seen from Fig. 6 that the problem of N rods on a line of length L is equivalent to the problem of N+1 rods on a circle of circumference C=L+a or radius

$$R = (L+a)/2\pi \tag{2.97}$$

Essentially, the edges of the first rod laid down (rod N+1) form the boundaries of the line segment 0<x<L, which we imagine to be wrapped around the circle as shown in Fig. 6. Thus, letting the angle θ_k locate the center of rod k relative to any chosen axis, we first locate rod N+1 by putting

$$\theta_{N+1} = 2\pi r_{N+1} \tag{2.98a}$$

where r_{N+1} is a random number from the uniform distribution in the unit interval. Then letting x measure the circumferential length from the leading edge of rod N+1 (x=0) to its trailing edge (x=C-a=L), we proceed

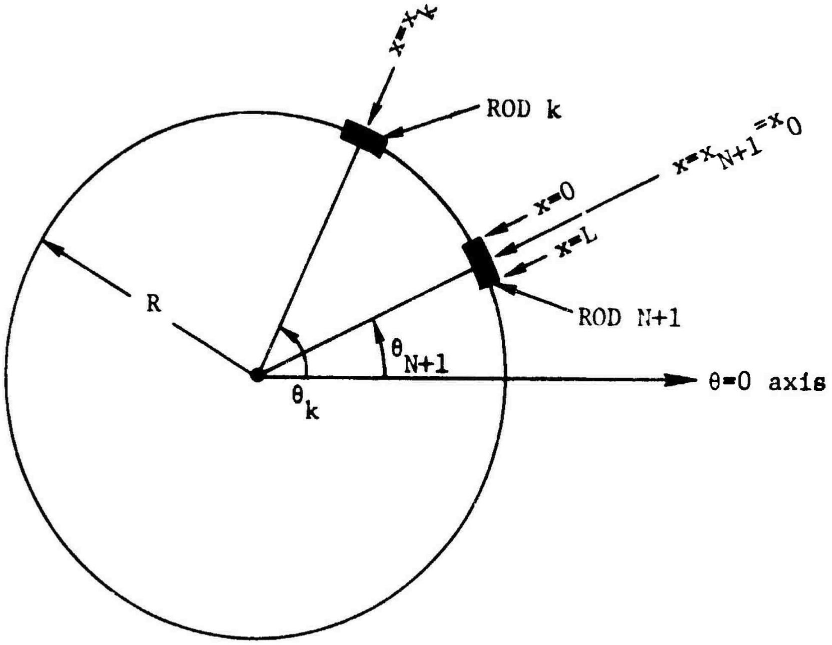

FIGURE 6. N+1 non-overlapping rods of equal length a on the circumference of a circle of radius $R=(L+a)/2\pi$.

to distribute the remaining N rods in $0<x<L$ as before. Thus, the angular location of rod k is [see Fig. 6]

$$\theta_k = \theta_{N+1} + (a/2 + x_k)/R, \quad k=N,N-1,\ldots,1 \tag{2.98b}$$

where x_k is generated according to (2.94).

As a final comment, and in anticipation of the development to be presented in the next chapter, we might point out that the generating algorithm developed in this section has potential applicability in the calculation of the thermodynamic properties of a "one-dimensional gas of hard-core rods". This is the one-dimensional analogue of a (non-ideal) gas composed of spherically symmetric molecules, which are assumed to have the property that an infinite repulsive force develops between any two molecules when the distance between their centers becomes equal to some fixed value $a>0$; a is called the "hard core diameter" of the molecules. Suppose a one-dimensional gas of N rods with hard-core diameter a is enclosed in the "volume" $0\leqslant x\leqslant L$, and is allowed to come to thermodynamic equilibrium with its surroundings at some absolute temperature T. The theory of statistical mechanics then tells us that the equilibrium value of any dynamical quantity f which depends only on the positions of the rods may be calculated as

$$\langle f\rangle_{eq} = \frac{\int_\Omega f(\vec{x})\exp[-U(\vec{x})/kT]d\vec{x}}{\int_\Omega \exp[-U(\vec{x})/kT]d\vec{x}} \tag{2.99}$$

Here, $\vec{x}=(x_1,x_2,\ldots,x_N)$ denotes a point in the allowable configuration space of the gas, which evidently is precisely the volume Ω in (2.77); $f(\vec{x})$ is the value of the dynamical quantity f for the configuration $\vec{x}$; $U(\vec{x})$ is the total potential energy of the rods for the configuration $\vec{x}$; and k is Boltzmann's constant. Analytical evaluations of the integrals in (2.99) have been successfully carried out for certain special forms of $f(\vec{x})$ and $U(\vec{x})$. However, as we shall see in the next chapter, the availability of an efficient algorithm for generating points $\vec{x}_i$ randomly and uniformly inside Ω opens up the possibility of numerically estimating these integrals by Monte Carlo methods for rather general forms of $f(\vec{x})$ and $U(\vec{x})$. A drawback of the Monte Carlo approach, as compared to a purely analytical approach, is that the dependence of $\langle f \rangle_{eq}$ on such external parameters as T, N and L can be inferred only by making separate calculations at specified values of these parameters; in addition, limitations on computation time will clearly place an upper limit on the size of N. Nevertheless, in cases where an analytical calculation simply cannot be effected, a series of Monte Carlo calculations, however restricted, may yield useful and otherwise unobtainable information.

Chapter 3

MONTE CARLO ESTIMATION OF INTEGRALS

3-1. Averages and Integrals

Let $f(\vec{x})$ denote any bounded function defined in the n-dimensional space of the variable $\vec{x}=(x^{(1)},x^{(2)},\ldots,x^{(n)})$, and let $P(\vec{x})$ denote a probability density function defined in this same space. We define "the average of $f(\vec{x})$ taken over the set of random points $\{\vec{x}_i\}$ distributed according to the density function $P(\vec{x})$" by[+]

$$\langle f:P\rangle \equiv \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} f(\vec{x}_i) \qquad (3.1)$$

whenever the limit exists. It is important to recognize that the average of $f(\vec{x})$ depends not only on the function f but also on the density function P which defines the set of random points $\{\vec{x}_i\}$ over which the average is taken. Thus, if we denoted the average of f by simply $\langle f\rangle$ our notation would be ambiguous and incomplete.

In the limit of sufficiently large N we may expect that $P(\vec{x})d\vec{x}$ accurately represents the fraction of the random points $\vec{x}_1,\vec{x}_2,\ldots,\vec{x}_N$ being summed over in (3.1) which falls inside the infinitesimal region $d\vec{x}$ centered

[+] Throughout this paper, the colon ":" within a mathematical expression can be verbalized as "with respect to". Thus, for example, $\langle f:P\rangle$ is read "the average of $f(\vec{x})$ with respect to the density function $P(\vec{x})$".

at the point $\vec{x}$. Therefore, in the limit of large N the contribution to the sum on the right side of (3.1) due to points $\vec{x}_i$ which lie inside $d\vec{x}$ at $\vec{x}$ is just $f(\vec{x})\times NP(\vec{x})d\vec{x}$. The sum in (3.1) can thus be calculated by summing (integrating) this quantity over all such infinitesimal regions in $\vec{x}$-space:

$$(\lim N\to\infty): \quad \sum_{i=1}^{N} f(\vec{x}_i) = \int_{\infty} f(\vec{x})NP(\vec{x})d\vec{x} \tag{3.2}$$

Dividing through by N and comparing with (3.1) gives the important result

$$\int_{\infty} f(\vec{x})P(\vec{x})d\vec{x} = \langle f:P\rangle \tag{3.3}$$

This result says that the integral of the quantity $f(\vec{x})P(\vec{x})$ over all $\vec{x}$-space is equal to the average of $f(\vec{x})$ taken with respect to the set of random points $\{\vec{x}_i\}$ distributed according to the density function $P(\vec{x})$. It forms the basis for the "Monte Carlo method" of evaluating a precisely defined definite integral as an average taken over a suitable set of random points. In particular, suppose the density function $P(\vec{x})$ defines the set of random points $\{\vec{x}_i\}$ distributed <u>uniformly</u> over some given finite region Ω of $\vec{x}$-space; i.e., suppose $P(\vec{x})$ is given by

$$P_\Omega(\vec{x}) \equiv \begin{cases} |\Omega|^{-1}, & \text{if } \vec{x}\varepsilon\Omega \\ 0 \ , & \text{if } \vec{x}\notin\Omega \end{cases} \tag{3.4}$$

where $|\Omega|$ is the volume of Ω. Then we have from (3.3) and (3.4)

$$\langle f:P_\Omega\rangle = \int_{\infty} f(\vec{x})P_\Omega(\vec{x})d\vec{x} = \int_{\Omega} f(\vec{x})|\Omega|^{-1}d\vec{x}$$

whence

$$\int_{\Omega} f(\vec{x})d\vec{x} = |\Omega|\left\langle f:P_{\Omega}\right\rangle \tag{3.5}$$

Thus, the integral of $f(\vec{x})$ over the region Ω is equal to the volume of Ω times the average of $f(\vec{x})$ with respect to the set of random points $\{\vec{x}_i\}$ distributed uniformly over Ω. This result is evidently a special case of (3.3).

If we could actually calculate the average $\langle f:P\rangle$ as defined in (3.1), then (3.3) would provide us with our Monte Carlo method for calculating definite integrals, and our story would be finished. In practice, of course, all we can really do is calculate quantities such as

$$\langle f:P\rangle_N \equiv \frac{1}{N}\sum_{i=1}^{N} f(\vec{x}_i) \tag{3.6}$$

for some finite N. Since by definition $\langle f:P\rangle_N \rightarrow \langle f:P\rangle$ as $N\rightarrow\infty$, then we may expect that if N is taken "fairly large"

$$\langle f:P\rangle_N \simeq \langle f:P\rangle, \tag{3.7}$$

so that the integral on the left of (3.3) is given approximately by the N-term average on the right of (3.6). But how good is this approximation? Clearly the approximation is worthless from a practical point of view unless we can give a meaningful estimate of its associated uncertainty. This important matter is the topic of the next section.

3-2. Fluctuations and Uncertainties

The short derivation of (3.3) given in the preceding section is useful for conveying an intuitive feeling for the central idea of the Monte Carlo method, but it is of little practical use in view of the fact that we can calculate only $\langle f{:}P\rangle_N$ in (3.6) and not $\langle f{:}P\rangle$ in (3.1). In fact, the goodness of the approximation in (3.7) can be obtained only by resorting to a famous but difficult-to-prove result in probability theory called the Central Limit Theorem. In order to state this theorem and to see how it applies to our problem here, it is first necessary to introduce a few concepts and definitions from probability theory.

Suppose we have a set of random real numbers $\{y_i\}$ distributed according to some density function $\tilde{P}(y)$. If $g(y)$ is any function of y, then, as we have seen in (3.1)-(3.3), the average of $g(y)$ over the set $\{y_i\}$ is given by

$$\langle g{:}\tilde{P}\rangle \equiv \lim_{K\to\infty} \frac{1}{K}\sum_{j=1}^{K} g(y_j) = \int_{-\infty}^{\infty} g(y)\tilde{P}(y)dy \tag{3.8}$$

In particular, we define the <u>mean</u> m and <u>variance</u> σ^2 of the set $\{y_i\}$ to be the averages over $\{y_i\}$ of the respective functions $g(y)=y$ and $g(y)=(y-m)^2$, assuming as we do that these averages exist:

$$m \equiv \langle y{:}\tilde{P}\rangle \equiv \lim_{K\to\infty} \frac{1}{K}\sum_{j=1}^{K} y_j = \int_{-\infty}^{\infty} y\tilde{P}(y)dy \tag{3.9}$$

$$\sigma^2 \equiv \langle (y-m)^2{:}\tilde{P}\rangle \equiv \lim_{K\to\infty} \frac{1}{K}\sum_{j=1}^{K} (y_j - m)^2 = \int_{-\infty}^{\infty} (y-m)^2\tilde{P}(y)dy \tag{3.10a}$$

By expanding the square in (3.10a) it is easy to show that the variance can also be written as the difference between the average of the square of y and the square of the average of y:

$$\sigma^2 = \int_{-\infty}^{\infty} y^2 \tilde{P}(y)dy - m^2 = \left\langle y^2 : \tilde{P} \right\rangle - \left\langle y : \tilde{P} \right\rangle^2 \qquad (3.10b)$$

The square root of the variance, namely σ, is called the root-mean-square (or rms or standard) deviation of the set $\{y_i\}$. If the graph of $\tilde{P}(y)$-versus-y consists mainly of a single hump, as is shown in Fig. 7, then (3.9) and (3.10a) imply that m and σ characterize respectively the "center" and "width" of the graph; roughly speaking, we may "reasonably expect" that a randomly drawn number from the set $\{y_i\}$ will lie somewhere between $m-\alpha\sigma$ and $m+\alpha\sigma$, where α is of order unity.

Now, suppose we construct from the set $\{y_i\}$ another set of random numbers $\{y_i^{(N)}\}$ according to the rule

$$y_i^{(N)} = (y_1 + y_2 + \ldots + y_N)/N \qquad (3.11)$$

That is, each element of the set $\{y_i^{(N)}\}$ is the simple average of N randomly drawn elements of the set $\{y_i\}$. We now ask, what can be said concerning the distribution of the set $\{y_i^{(N)}\}$?

In Appendix C [cf. (C.14)-(C.16)] we derive a partial answer to this question: the mean and rms deviation of the set $\{y_i^{(N)}\}$ are equal to m and $\sigma/\sqrt{N}$ respectively, where m and σ are as before the mean and rms deviation of the original set $\{y_i\}$. However, for Monte Carlo purposes it is necessary to have a somewhat more detailed knowledge of the distribution of the set $\{y_i^{(N)}\}$; specifically, we would like to be able to translate the rms deviation $\sigma/\sqrt{N}$ of this set into a specification of numerical confidence limits. For this we must resort to the Central Limit Theorem [see p. 244 of Ref. 8],

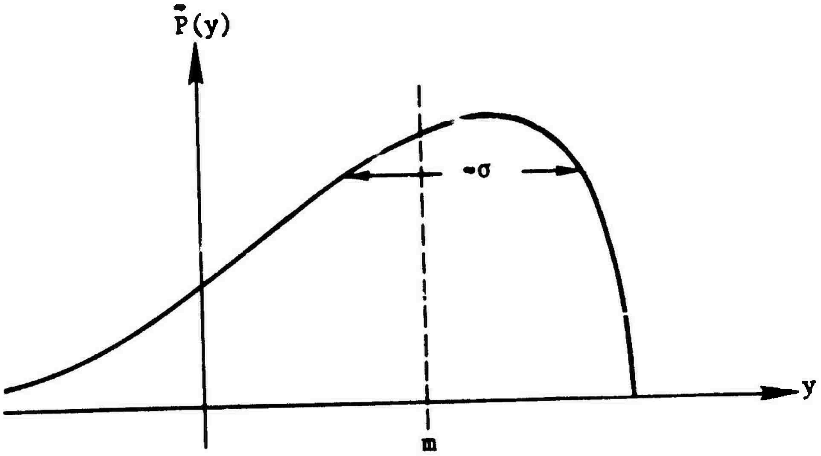

FIGURE 7. Illustrating the geometrical significance of the mean m and rms deviation σ of a set of random numbers $\{y_i\}$ with density function $\tilde{P}(y)$.

which says in effect that *in the limit of large* N the set $\{y_i^{(N)}\}$ becomes a *Gaussian* (or "normal") distribution, with (as just asserted) mean m and rms deviation $\sigma/\sqrt{N}$. More precisely, the Central Limit Theorem asserts that

$$\lim_{N\to\infty} \mathrm{Prob}\left|\,|y_i^{(N)} - m| \leqslant \alpha\frac{\sigma}{\sqrt{N}}\right| = \sqrt{\frac{1}{2\pi}}\int_{-\alpha}^{\alpha}\exp(-t^2/2)dt$$

$$\simeq \begin{cases} 0.683 & \text{for } \alpha=1 \\ 0.955 & \text{for } \alpha=2 \\ 0.997 & \text{for } \alpha=3 \end{cases} \qquad (3.12)$$

Thus, for example, *provided N is sufficiently large* the probability that the average of N randomly drawn elements from the set $\{y_i\}$ will lie between $m-2\sigma/\sqrt{N}$ and $m+2\sigma/\sqrt{N}$ is 0.955, where m and σ are respectively the mean and rms deviation of $\{y_i\}$. The power of this result is that it holds irrespective of the form of the density function $\tilde{P}(y)$ of $\{y_i\}$. However, the Central Limit Theorem also has a notable limitation: it does *not* specify how large N must be in order that the *lim* symbol in (3.12) can be ignored for practical purposes. Presumably, the rate of approach to the limit in (3.12) will depend in some complicated way upon specific form of the density function $\tilde{P}(y)$, but the Central Limit Theorem gives us no information in this regard.

Now let us see how the foregoing results allow us to quantitatively estimate the uncertainty associated with the crucial approximation (3.7). Essentially we have taken a set of random *points* $\{\vec{x}_i\}$ distributed in n-dimensional space according to the density function $P(\vec{x})$, and we have constructed a set of random *numbers* $\{y_i\}$ by subjecting $\{\vec{x}_i\}$ to the transformation

$$y = f(\vec{x}) \tag{3.13}$$

This set of random numbers $\{y_i\}$ will have some density function $\tilde{P}(y)$, the form of which is completely determined by the two functions $P(\vec{x})$ and $f(\vec{x})$. In practice, it is virtually impossible to calculate the shape of $\tilde{P}(y)$ analytically from $P(\vec{x})$ and $f(\vec{x})$; just to illustrate what would be involved in such a calculation, we sketch in Fig. 8c the density function $\tilde{P}(y)$ which would result in the case where $\vec{x}$ is one-dimensional and the functions $P(\vec{x})$ and $f(\vec{x})$ are those sketched in Figs. 8a and 8b respectively[†]. But in

[†] Given a set of random points $\{x_i\}$ distributed according to the density function $P(x)$ in Fig. 8a, then the transformation $y = f(x)$ in Fig. 8b produces a set of random numbers $\{y_i\}$ whose density function $\tilde{P}(y)$ is obtained from the general rule $\tilde{P}(y)dy=P(x)dx$—i.e., $\tilde{P}(y)=P(x)/|f'(x)|$ where $x=f^{-1}(y)$ with proper account being taken of the multi-valuedness of the inverse function $f^{-1}(y)$. Thus, for Figs. 8a and 8b we have

$$\tilde{P}(y) = \tilde{P}_1(y) + \tilde{P}_2(v)$$

where $$\tilde{P}_1(y) = \begin{cases} P(x)/|f'(x)|, & \text{for} \quad x<\xi \\ 0 \quad , & \text{for} \quad x>\xi \end{cases}$$

and $$\tilde{P}_2(y) = \begin{cases} 0 \quad , & \text{for} \quad x<\xi \\ P(x)/|f'(x)|, & \text{for} \quad x>\xi \end{cases}$$

The resultant $\tilde{P}(y)$ is sketched in Fig. 8c. The jump discontinuity in $\tilde{P}(y)$ at $y=B$ is due to the fact that the interval $B<y<C$ is fully populated by y_i's coming, not only from x_i's in the interval $a<x<\xi$, but also from x_i's in the interval $\xi<x<\xi'$. The infinite peak in $\tilde{P}(y)$ at $y=C$ is due to the fact that $f'(\xi)=0$. The hump in $\tilde{P}(y)$ near $y=A$ is a reflection of the hump in $P(x)$ near $x=b$.

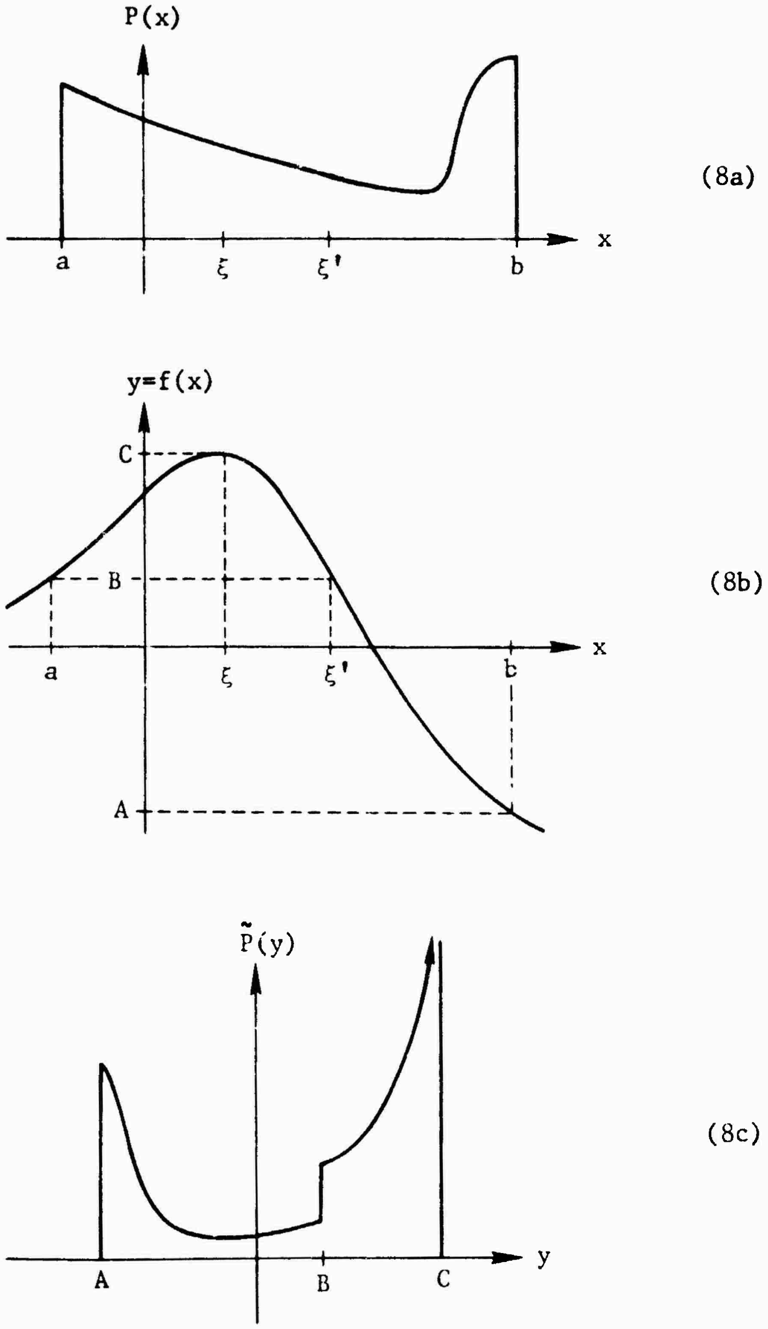

FIGURE 8. Illustrating how the density function $\tilde{P}(y)$ of the values $y=f(\overrightarrow{x})$ is determined by the density function $P(\overrightarrow{x})$ and the function $f(\overrightarrow{x})$, in a simple case in which $\overrightarrow{x}$ is one-dimensional. (See footnote on p. 83.)

any case, it is clear that the mean m of the set $\{y_i\}$—i.e., the "center" of the curve of $\tilde{P}(y)$-versus-y—coincides with the average of $f(\vec{x})$ with respect to the set $\{\vec{x}_i\}$:

$$m \equiv \langle y:\tilde{P}\rangle \equiv \lim_{K\to\infty} \frac{1}{K} \sum_{j=1}^{K} y_j = \lim_{K\to\infty} \frac{1}{K} \sum_{j=1}^{K} f(\vec{x}_j) \equiv \langle f:P\rangle \tag{3.14a}$$

Similarly, the rms deviation σ of the set $\{y_i\}$—i.e., the "width" of the curve of $\tilde{P}(y)$-versus-y—is given by

$$\sigma \equiv \sqrt{\langle y^2:\tilde{P}\rangle - \langle y:\tilde{P}\rangle^2} = \sqrt{\langle f^2:P\rangle - \langle f:P\rangle^2} \tag{3.14b}$$

And finally, the average $y_i^{(N)}$ of N random elements of $\{y_i\}$ is seen to coincide with the calculated quantity $\langle f:P\rangle_N$:

$$y_i^{(N)} \equiv \frac{1}{N}\sum_{j=1}^{N} y_j = \frac{1}{N} \sum_{j=1}^{N} f(\vec{x}_j) \equiv \langle f:P\rangle_N \tag{3.14c}$$

In view of Eqs. (3.14) we thus see that the Central Limit Theorem in (3.12) implies that

$$\lim_{N\to\infty} \text{Prob}\left\{\left|\langle f:P\rangle_N - \langle f:P\rangle\right| \leqslant \alpha \frac{\sigma}{\sqrt{N}}\right\} = \sqrt{\frac{1}{2\pi}} \int_{-\alpha}^{\alpha} \exp(-t^2/2)dt \tag{3.15}$$

where $\quad \sigma = \sqrt{\langle f^2:P\rangle - \langle f:P\rangle^2} \quad (3.16)$

In essence, then, we may say that <u>for</u> N <u>sufficiently large</u> the approximation in (3.7) has a Gaussian rms uncertainty of $\sigma/\sqrt{N}$, where σ is given by (3.16). Of course, since we do not know $\langle f:P\rangle$ beforehand, much less $\langle f^2:P\rangle$, then it is clear that we do not know σ beforehand either. However, if N is sufficiently large so that the approximation $\langle f:P\rangle_N \simeq \langle f:P\rangle$ is good, then

then we may also put $\langle f^2:P\rangle_N \simeq \langle f^2:P\rangle$, and thereby take

$$\sigma \simeq \sqrt{\langle f^2:P\rangle_N - \langle f:P\rangle_N^2} \tag{3.17}$$

as a sufficiently accurate approximation to σ for the purposes of (3.15).

In summary, then, we have the following "Golden Rule of Monte Carlo": If $\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_N$ are N random points distributed according to the probability density function $P(\vec{x})$, and if for a given function $f(\vec{x})$ we put

$$\langle f:P\rangle_N \equiv \frac{1}{N}\sum_{i=1}^{N} f(\vec{x}_i) \quad \text{and} \quad \langle f^2:P\rangle_N \equiv \frac{1}{N}\sum_{i=1}^{N} f^2(\vec{x}_i) \tag{3.18}$$

then provided N is sufficiently large we have

$$\int_\infty f(\vec{x})P(\vec{x})d\vec{x} \simeq \langle f:P\rangle_N \pm \frac{\sqrt{\langle f^2:P\rangle_N - \langle f:P\rangle_N^2}}{\sqrt{N}} \tag{3.19}$$

The ± quantity in (3.19) is understood to correspond to a "65% confidence interval", and should thus be typical of the average spread between several independent evaluations of $\langle f:P\rangle_N$. If we double the ± uncertainty in (3.19) we obtain a "95% confidence interval", which is perhaps a more suitable uncertainty to use when asserting a value for the integral; and if we treble the ± uncertainty, we obtain an even more conservative "99% confidence interval". Note that in the limit $N\to\infty$ (3.19) indeed goes over into (3.3).

3-3. Operational Procedures

In what follows we shall assume that we are presented with the problem of integrating a bounded function $f(\vec{x})$ over a finite region Ω in n-dimensional $\vec{x}$-space. Now,

$$\int_{\Omega} f(\vec{x})d\vec{x} = |\Omega|\int_{\Omega} f(\vec{x})|\Omega|^{-1}d\vec{x} = |\Omega|\int_{\Omega} f(\vec{x})P_{\Omega}(\vec{x})d\vec{x}$$

or

$$\int_{\Omega} f(\vec{x})d\vec{x} = |\Omega|\int_{\infty} f(\vec{x})P_{\Omega}(\vec{x})d\vec{x} \tag{3.20}$$

where $|\Omega|$ is the volume of Ω and $P_{\Omega}(\vec{x})$ is the density function (3.4) defining the uniform distribution of random points inside Ω. Substituting (3.19) into the right side of (3.20) then yields the result

$$\int_{\Omega} f(\vec{x})d\vec{x} \simeq |\Omega|\left\{\langle f:P_{\Omega}\rangle_N \pm \frac{\sqrt{\langle f^2:P_{\Omega}\rangle_N-\langle f:P_{\Omega}\rangle_N}}{\sqrt{N}}\right\} \tag{3.21}$$

The procedure for numerically estimating the integral of a bounded function $f(\vec{x})$ over a finite region Ω of $\vec{x}$-space therefore consists of the following steps, which we state in the "iterative language" of linear computer programming:

1° Initialize $S_1=0$, $S_2=0$, $i=0$.

2° Generate a random point $\vec{x}$ from the uniform distribution inside Ω [i.e., according to the density function $P_{\Omega}(\vec{x})$ in (3.4)].

3° Evaluate $y=f(\vec{x})$ for the generated point $\vec{x}$.

4° Put $S_1=S_1+y$, $S_2=S_2+y^2$, $i=i+1$.

5° If $i<N$, where N is as large an integer as practical, return to step 2°; if $i=N$ go to step 6°.

6° Put $\langle f:P_{\Omega}\rangle_N=S_1/N$ and $\langle f^2:P_{\Omega}\rangle_N=S_2/N$, and obtain the Monte Carlo estimate of the integral from (3.21). Remember that the ± uncertainty in (3.21) represents a 65% confidence interval; in a final quotation it is usually best to double this uncertainty to obtain a 95% confidence interval.

(3.22)

The foregoing steps are schematized in the flow diagram of Fig. 9.

It will be observed that the above procedure *always* yields numerical results for the estimate and its associated uncertainty, even if N has not been chosen large enough to satisfy the limit requirement in (3.15) and the σ approximation in (3.17). It is important to realize that, *unless* N *is* *large* *enough* for these two approximations to be valid, then one may *not* assert that the calculated estimate is indeed accurate to within the calculated uncertainty with the numerical confidence limits prescribed in (3.12). Unfortunately, there is no way of telling from the results of a *single* Monte Carlo calculation whether N has been taken large enough to produce reliable results. One fairly convenient way of checking this important point is to proceed as follows: Choose a value of N such that one can afford to perform the Monte Carlo calculation *four* times, each time using a different set of N random points. If the *uncertainties* obtained in these four runs agree to within better than 10% (i.e., to within at least the first significant digit), and if further the four *estimates* each differ from their average by no more than roughly twice the uncertainty, then one usually can accept these results subject to the confidence limits prescribed in (3.12). One may then quote as the estimate of the integral the *average* of the estimates found in the four separate runs. Since this average evidently corresponds to a total of 4N points, then its one-standard deviation uncertainty will be exactly half that of the individual runs; thus, the *one*-standard deviation uncertainty found in the four individual runs may be quoted as the *two*-standard deviation uncertainty (giving 95%

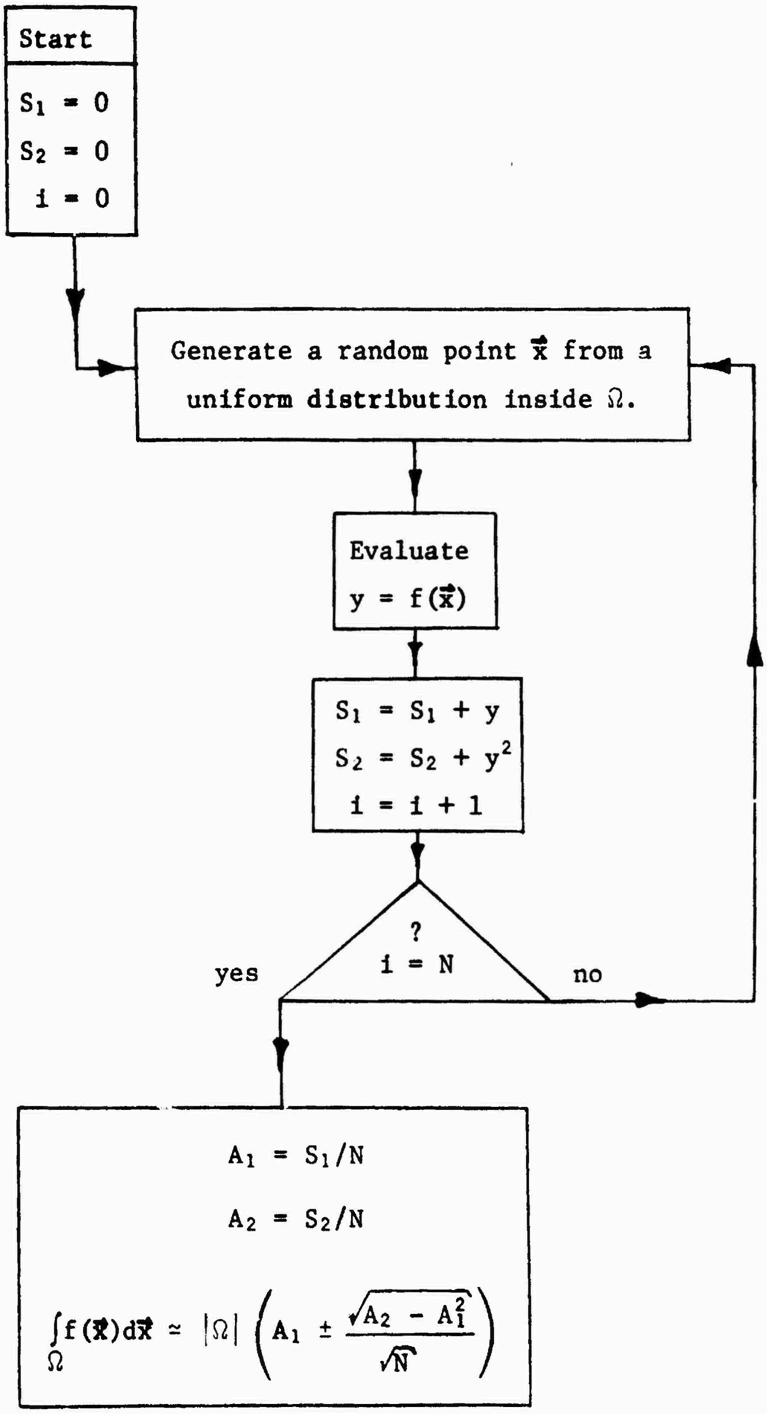
Start
$S_1 = 0$
$S_2 = 0$
$i = 0$
Generate a random point $\vec{x}$ from a uniform distribution inside Ω.
Evaluate $y = f(\vec{x})$
$S_1 = S_1 + y$
$S_2 = S_2 + y^2$
$i = i + 1$
$i \overset{?}{=} N$
yes
no
$A_1 = S_1/N$
$A_2 = S_2/N$
$\int_\Omega f(\vec{x})d\vec{x} \simeq |\Omega| \left(A_1 \pm \frac{\sqrt{A_2 - A_1^2}}{\sqrt{N}} \right)$

FIGURE 9. Flow diagram of the basic steps involved in the Monte Carlo evaluation of an integral.

confidence limits) for the average of the estimates found in the individual runs.†

If, on the other hand, the results of the four repeated Monte Carlo runs are not mutually compatible in the sense described above, then one should tentatively assume that N is not large enough to produce reliable results. If increasing N fails to improve the situation, then one should suspect the existence of either (i) a programming error, or (ii) a singularity in the integrand. The latter circumstance is of course intolerable; for, if $f(\vec{x})$ approaches infinity as $\vec{x}$ approaches some point $\vec{x}_c$ in Ω (but in such a way that $\int_\Omega f(\vec{x})d\vec{x}$ is nevertheless finite), then the Monte Carlo estimate of the integral can be arbitrarily large, depending solely upon how close one happens to come to $\vec{x}_c$ while picking points at random in Ω.

Our main point here, though, is that the entire Monte Carlo method is predicated on the assumption that the number of random points sampled is "sufficiently large"; therefore, one should never accept the numerical results of the computational procedure (3.22)[or Fig. 9] without being reasonably confident that N is indeed "sufficiently large". One way of checking this point, as suggested above, is to require consistency among the results of several repeated calculations. In the sequel, we shall always assume that this or some equivalent check has been performed.

† An additional advantage of splitting a 4N-point Monte Carlo run into four separate N-point runs is that it considerably reduces the possible loss which might result from a computer malfunction or control card error.

A common source of difficulty in executing the six-step procedure of (3.22) is generating a random point from a uniform distribution inside Ω (Step 2°). If Ω is a box-like region then the calculation of $|\Omega|$ and the generation of random points uniformly inside Ω is very easy [cf. (2.35]. However, it frequently happens that Ω is not a box-like region, and that we are not able to calculate $|\Omega|$; this will usually be the case if Ω is defined by a set of inequalities involving various functions of the components of $\vec{x}$ which are not "neatly ordered" as in (2.31a). In such a case the following procedure can sometimes be used. Find some box-like region Σ which completely encloses Ω, and define a new function $g(\vec{x})$ in Σ which coincides with f inside Ω but which vanishes outside Ω:

$$\left.\begin{array}{l} \Sigma \text{ (box-like) encloses } \Omega \\ g(\vec{x}) \equiv \begin{cases} f(\vec{x}), & \text{if } \vec{x}\epsilon\Omega \\ 0 \;\;, & \text{if } \vec{x}\not\epsilon\Omega \end{cases} \end{array}\right\} \tag{3.23}$$

Clearly, the integral of $g(\vec{x})$ over Σ is equal to the integral of $f(\vec{x})$ over Ω; thus, the procedure (3.22) can be carried out with f and Ω replaced by g and Σ respectively, and (3.21) becomes

$$\int_\Omega f(\vec{x})d\vec{x} \simeq |\Sigma| \left\{ \langle g:P_\Sigma\rangle_N \pm \frac{\sqrt{\langle g^2:P_\Sigma\rangle_N - \langle g:P_\Sigma\rangle_N^2}}{\sqrt{N}} \right\} \tag{3.24}$$

Of course, we should not expect this procedure to be very efficient if $|\Omega|<<|\Sigma|$, and for this reason we should always take Σ to be the smallest box enclosing Ω.

A frequently useful example of the foregoing procedure is the calculation of the volume $|\Omega|$ itself. Since

$$\int_{\Omega} 1 \cdot d\vec{x} = |\Omega|$$

then we have from (3.23) and (3.24)

$$|\Omega| \simeq |\Sigma| \left\{ \langle g:P_{\Sigma} \rangle_N \pm \frac{\sqrt{\langle g^2:P_{\Sigma} \rangle_N - \langle g:P_{\Sigma} \rangle_N^2}}{\sqrt{N}} \right\} \tag{3.25a}$$

where $$g(\vec{x}) \equiv \begin{cases} 1, & \text{if } \vec{x} \in \Omega \\ 0, & \text{if } \vec{x} \notin \Omega \end{cases} \tag{3.25b}$$

Now suppose that, of the N points generated randomly and uniformly inside Σ, M of them are found to fall inside Ω. It follows from (3.25b) that the sums S_1 and S_2 computed in accordance with Step 4° of (3.22) will both equal M, so that in Step 6° we will find

$$\langle g:P_{\Sigma} \rangle_N = \langle g^2:P_{\Sigma} \rangle_N = M/N \tag{3.26}$$

Substituting into (3.25a) yields the result

$$|\Omega| \simeq \frac{M}{N} |\Sigma| \left[1 \pm \sqrt{\frac{N - M}{N \cdot M}} \right] \tag{3.27}$$

where, again, of N uniformly distributed random points inside Σ, M are found to lie inside Ω. According to (3.27), |Ω| is given approximately by (M/N)|Σ|, just as we expect. But (3.27) also provides us with an estimate of the uncertainty in this approximation. Evidently, the relative uncertainty is equal to $1/\sqrt{N}$ times the square root of the ratio of the number of "miss points" N-M to the number of "hit points" M; therefore, as mentioned earlier, the uncertainty will be large if M<<N, or |Ω|<<|Σ|.

In fact, by putting $M/N \simeq |\Omega|/|\Sigma|$ in the relative uncertainty term in (3.27), it is easy to show that if $|\Omega|$ is n(>1) orders of magnitude smaller than $|\Sigma|$, then N will have to be of order n+2 to obtain an estimate of $|\Omega|$ having a 10% relative uncertainty, and of order n+4 to obtain an estimate having a 1% relative uncertainty.

Returning to the problem of integrating a general function $f(\vec{x})$ over a non-box region Ω, suppose the volume $|\Omega|$ is known exactly, but that it is nevertheless not practical to generate uniformly distributed random points $\vec{x}$ inside Ω by the direct (inversion) method. In such a case one could of course still use the box-method described in connection with (3.23) and (3.24); alternatively, one could use (3.21) in conjunction with the rejection method for generating random points uniformly inside Ω. Thus, using the same enclosing box Σ as in (3.23) and (3.24), suppose that, of N random points generated uniformly inside the box Σ, M are found to fall inside Ω. Then these latter M points can be used to calculate the quantities $\langle f:P_\Omega\rangle_M$ and $\langle f^2:P_\Omega\rangle_M$, and we may put in accordance with (3.21)

$$\int_\Omega f(\vec{x})d\vec{x} \simeq |\Omega| \left\{ \langle f:P_\Omega\rangle_M \pm \frac{\gamma\sqrt{\langle f^2:P_\Omega\rangle_M - \langle f:P_\Omega\rangle_M^2}}{\sqrt{M}} \right\} \tag{3.28}$$

It is interesting to compare the approach of (3.28) with that of (3.24). Given the precise values of both volumes $|\Omega|$ and $|\Sigma|$, it should be obvious that calculating according to (3.28) involves exactly the same amount of work as calculating according to (3.24); indeed, it is clear that the sums S_1 and S_2 calculated in Step 4° of (3.22) will be the same for (3.28) as for (3.24), because the N-M points $\vec{x}_i$ not used in these sums for (3.28) contribute zero to these sums for (3.24). In terms of these common sums S_1 and S_2, (3.24) and (3.28) can be written respectively as

$$\int_\Omega f(\vec{x})d\vec{x} \simeq |\Sigma|\frac{S_1}{N}\left(1 \pm \sqrt{\frac{S_2}{S_1^2} - \frac{1}{N}}\right) \tag{3.29a}$$

and

$$\int_\Omega f(\vec{x})d\vec{x} \simeq |\Omega|\frac{S_1}{M}\left(1 \pm \sqrt{\frac{S_2}{S_1^2} - \frac{1}{M}}\right) \tag{3.29b}$$

Since in the limit $N\to\infty$, we expect $|\Omega|/|\Sigma| = M/N$, then the averages or central values in (3.29a) and (3.29b) are indeed consistent. But as M is always less than N, then the relative uncertainty in (3.29b) is always less than that in (3.29a). If $f(\vec{x})$ is approximately constant in Ω we will have $S_1^2 \simeq MS_2$, so that the relative uncertainty in (3.29a) will be approximately that given in (3.27) whereas the relative uncertainty in (3.29b) will be approximately zero. Thus, if $|\Omega|$ is known exactly, then it is usually a bit more efficient to make use of this knowledge and proceed via (3.28) rather than via (3.24).

If the volume of Ω is very much smaller than the volume of the smallest box Σ which encloses Ω, then clearly neither (3.28) nor (2.34) will be satisfactory. In such a case one should endeavor to calculate $|\Omega|$ analytically, and in the process also calculate the distribution functions $F(x^{(1)})$, $F(x^{(2)}|x^{(1)}), \ldots, F(x^{(n)}|x^{(1)}, \ldots, x^{(n-1)})$ which allow one to use the generalized inversion method for directly generating random points $\vec{x}_i$ uniformly inside Ω

[cf. (2.34) and (2.29)]. Alternatively, one can try to find some transformation of variables $\vec{x}\to\vec{x}'$ which carries the region Ω into a more suitably shaped region Ω'. Given such a transformation (take n=3 for concreteness),

$$\left.\begin{aligned} x' &= x'(x,y,z) \\ y' &= y'(x,y,z) \\ z' &= z'(x,y,z) \end{aligned}\right\} \qquad (3.30a)$$

one could then express the given integral as an integral over Ω' according to [cf. (B.9)]

$$\iiint_{\Omega} f(x,y,z)dxdydz = \iiint_{\Omega'} f(x,y,z)\left|\frac{\partial(x,y,z)}{\partial(x',y',z')}\right| dx'dy'dz'$$

$$\equiv \iiint_{\Omega'} h(x',y',z')dx'dy'dz' \qquad (3.30b)$$

Here, the last step is carried out after solving (3.30a) for x, y and z in terms of x', y' and z'. It is interesting to note that the inversion method for generating random points uniformly inside Ω, in which one calculates the one-dimensional distribution functions $F(x)$, $F(y|x)$, $F(z|x,y)$ for the uniform density function $P_{\Omega}(x,y,z)$ and puts

$$\left.\begin{aligned} r_1 &= F(x) \\ r_2 &= F(y|x) \\ r_3 &= F(z|x,y) \end{aligned}\right\} \qquad (3.31a)$$

can also be regarded as a transformation of variables from xyz-space to $r_1r_2r_3$-space. This transformation has the special properties that (i) the region Ω in xyz-space is transformed into the unit cube in $r_1r_2r_3$-space, and (ii) the Jacobian of the transformation is simply [cf. (2.30) and (2.32)]

$$\frac{\partial(x,y,z)}{\partial(r_1,r_2,r_3)} = \frac{1}{P_\Omega(x,y,z)} = \frac{1}{|\Omega|^{-1}} = |\Omega| \qquad (3.31b)$$

There are, of course, many transformations which will carry Ω into a unit cube, but (3.31a) is the only one of these that has a constant Jacobian. Another transformation to the unit cube which can be applied whenever Ω is specified as in (2.31a) is the simple "linear stretching" transformation

$$\left.\begin{aligned} r_1 &= [x - a_1]/[b_1 - a_1] \\ r_2 &= [y - a_2(x)]/[b_2(x) - a_2(x)] \\ r_3 &= [z - a_3(x,y)]/[b_3(x,y) - a_3(x,y)] \end{aligned}\right\} \qquad (3.32a)$$

the Jacobian of which is easily calculated as

$$\frac{\partial(x,y,z)}{\partial(r_1,r_2,r_3)} = [b_1 - a_1][b_2(x) - a_2(x)][b_3(x,y) - a_3(x,y)] \qquad (3.32b)$$

Whether (3.31a) or (3.32a) is the better transformation to the unit cube depends on the form of the integrand, as we shall see more cl arly in our discussion of "importance sampling" in Secs. 4-5 and 4-6.

If, in the case where the volume of Ω is extremely small compared to the volume of the smallest box enclosing Ω, all efforts at inversion generating and variables transformation fail, then one may simply be forced to conclude that the conventional Monte Carlo method is not applicable.

An integral which has either an unbounded integrand or an infinite integrating region (or both) evidently presents a special problem. Usually the best procedure is to try to find a transformation of variables [cf. (3.30)] which is such that the transformed integrand (i.e., the old integrand times the Jacobian of the transformation) is bounded, and the transformed integrating region is finite. Then the methods outlined above may be applied. If the given integral truly exists, then many such transformations to a bounded integrand and a finite integrating region exist, but whether or not one of these can actually be found is another matter. In certain cases where the only difficulty is an infinite integrating region, we can often just use the more general Monte Carlo formula (3.19). For example, suppose we have to evaluate the one-dimensional integral

$$I = \int_0^\infty f(x)dx \qquad (3.33a)$$

where f(x) can be written in the form

$$f(x) = h(x)x^n \exp(-ax^2) \qquad (3.33b)$$

with n a non-negative integer, a>0, and h(x) bounded for 0<x<∞. Using the density function P(x;n,a) defined in (2.51), we may evidently write I as

$$I = A^{-1}(n,a)\int_0^\infty h(x)P(x;n,a)dx \qquad (3.34)$$

Since, as discussed in detail in Sec. 2-9, we know how to generate random numbers x_i in $0<x<\infty$ according to the density function $P(x;n,a)$, then we may evidently proceed according to (3.19), wherein the volume of the integrating region never enters explicitly. Note that if we use the straight inversion method to generate random points x_i according to $P(x;n,a)$—i.e., if we proceed by actually inverting

$$r = F(x;n,a) \qquad (3.35a)$$

for a given $r=r_i$ where $F(x;n,a)$ is the distribution function in (2.52)—then we are in effect making a change of variable $x\to r$: For, by (3.35a)

$$dr = F'(x;n,a)dx = P(x;n,a)dx \qquad (3.35b)$$

so that (3.34) can be transformed to

$$I = A^{-1}(n,a)\int_0^1 h(x)dr \qquad (3.36)$$

In (3.36) x is now understood to be the function of r obtained by inverting (3.35a). As discussed in connection with (2.52), this inversion must be done numerically in all cases except for n=1.

3-4. Combining Results of Monte Carlo Calculations

It frequently happens that we are interested in the sum of two integrals, each of which has been computed independently in separate Monte Carlo calculations. Consider, for example, the two integrals

$$I_k = \int_{\Omega_k} f_k(\vec{x}_k)d\vec{x}_k, \quad k = 1 \text{ and } 2 \qquad (3.37)$$

where we allow even for the possibility that Ω_1 and Ω_2 may be of different dimensionalities. If two independent Monte Carlo calculations have yielded the results [cf. (3.21)]

$$I_k \simeq \tilde{I}_k \pm \Delta_k, \quad k = 1 \text{ and } 2 \tag{3.38}$$

then, as we shall show below, we may assert that

$$I_1 + I_2 \simeq (\tilde{I}_1 + \tilde{I}_2) \pm \sqrt{\Delta_1^2 + \Delta_2^2} \tag{3.39}$$

Note in particular that, although the estimate of the sum is the sum of the estimates, the uncertainty in the sum is always somewhat less than the sum of the uncertainties.

To show that (3.39) is true we assume that, for k=1 and 2, $\tilde{I}_k$ and Δ_k have been calculated in the usual Monte Carlo way [cf. (3.21)] using a "sufficiently large" number of uniformly distributed random points in Ω_k. Then $\tilde{I}_k$ is in fact a particular element of a set of random numbers $\{\tilde{I}_{k,i}\}$ which has a mean I_k and a variance Δ_k^2. Now clearly $\tilde{I}_1 + \tilde{I}_2$ is a particular element of a set of random numbers $\{\tilde{I}_{1,i} + \tilde{I}_{2,i}\}$ which is formed by adding pairs of independently drawn random numbers from the two sets $\{\tilde{I}_{1,i}\}$ and $\{\tilde{I}_{2,i}\}$. The question is, what are the mean and variance of the set $\{\tilde{I}_{1,i} + \tilde{I}_{2,i}\}$? The answer is provided by a well-known theorem of statistics, proved in Appendix C, which says that the mean and variance of the sum of two statistically independent sets of random numbers are equal respectively to the sums of the means and variances of the two sets. Thus, the set of random numbers $\{\tilde{I}_{1,i} + \tilde{I}_{2,i}\}$ has mean $I_1 + I_2$ and variance $\Delta_1^2 + \Delta_2^2$.

These facts evidently enable us to make the statement (3.39).

Another frequently encountered situation is the following: We are presented with an integral I of the form

$$I = \int_{\Omega} f(\vec{x})d\vec{x} \tag{3.40a}$$

where $f(\vec{x})$ is the sum

$$f(\vec{x}) = f_1(\vec{x}) + f_2(\vec{x}) \tag{3.40b}$$

It is desired to calculate by Monte Carlo methods not only the integral I but also the two integrals

$$I_k \equiv \int_{\Omega} f_k(\vec{x})d\vec{x}, \quad k = 1 \text{ and } 2 \tag{3.41}$$

which evidently constitute I according to

$$I = I_1 + I_2 \tag{3.42}$$

Letting $\{\vec{x}_i\}$ denote the set of random points distributed uniformly over Ω [i.e., according to the density function $P_{\Omega}(\vec{x})$], then the set of random numbers $\{f_k(\vec{x}_i)\}$ has mean m_k and variance σ_k^2 given by

$$m_k = \langle f_k : P_{\Omega} \rangle$$

and

$$\sigma_k^2 = \langle f_k^2 : P_{\Omega} \rangle - \langle f_k : P_{\Omega} \rangle^2$$

If we use N random points $\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_N$ from $\{\vec{x}_i\}$ to compute the N-term average $\langle f_k:P_\Omega\rangle_N$ [cf. (3.6)], then if N is sufficiently large we can put

$$I_k \simeq |\Omega|\left(\langle f_k:P_\Omega\rangle_N \pm \frac{\sigma_k}{\sqrt{N}}\right), \quad k = 1 \text{ and } 2 \tag{3.43}$$

Now, if we use one set of N points from $\{\vec{x}_i\}$ to calculate $\langle f_1:P_\Omega\rangle_N$ and a different set of N points to calculate $\langle f_2:P_\Omega\rangle_N$, then these two estimates will be statistically independent; hence, we may invoke (3.39) and obtain for I the result

$$I \simeq |\Omega| \left(\langle f_1:P_\Omega\rangle_N + \langle f_2:P_\Omega\rangle_N \pm \frac{\sqrt{\sigma_1^2+\sigma_2^2}}{\sqrt{N}}\right) \tag{3.44}$$

Suppose, however, that we calculate $\langle f_1:P_\Omega\rangle_N$ and $\langle f_2:P_\Omega\rangle_N$ in (3.43) using the same set of N points from $\{\vec{x}_i\}$. This would obviously be desirable from the standpoint of calculating only I_1 and I_2, since it would require generating only N instead of 2N random points $\vec{x}_i$. Of course, if this were done then the random numbers $\langle f_1:P_\Omega\rangle_N$ and $\langle f_2:P_\Omega\rangle_N$ would not be statistically independent, and we could therefore not assert the result (3.44) for I. However, as we calculate $\langle f_1:P_\Omega\rangle_N$ and $\langle f_2:P_\Omega\rangle_N$ "in parallel" (i.e., using the same set of N random points), it would clearly require very little effort to also calculate, using (3.40b), the quantity $\langle f:P_\Omega\rangle_N$. Then instead of (3.44) we could assert for I the usual Monte Carlo result

$$I \simeq |\Omega|\left(\langle f:P_\Omega\rangle_N \pm \frac{\sigma}{\sqrt{N}}\right) \tag{3.45}$$

where

$$\sigma^2 = \langle f^2:P_\Omega\rangle - \langle f:P_\Omega\rangle^2$$

is the variance of the set $\{f(\vec{x}_i)\} \equiv \{f_1(\vec{x}_i)+f_2(\vec{x}_i)\}$. We note that, because of (3.40b), the quantity $\langle f:P_\Omega\rangle_N$ in (3.45) could also be written as the sum of $\langle f_1:P_\Omega\rangle_N$ and $\langle f_2:P_\Omega\rangle_N$, just as in (3.44). In other words, the estimate

$$I \simeq |\Omega|[\langle f_1:P_\Omega\rangle_N + \langle f_2:P_\Omega\rangle_N]$$

is valid regardless of whether $\langle f_1:P_\Omega\rangle_N$ and $\langle f_2:P_\Omega\rangle_N$ are computed using the same set or different sets of N random points from $\{\vec{x}_i\}$. However, the uncertainties in these estimates are not in general the same, because σ^2 in (3.45) is not generally equal to $\sigma_1^2+\sigma_2^2$ in (3.44).

In summary, then, we may calculate I_1,I_2 and I in either of two ways: One way is to first make separate Monte Carlo calculations of I_1 and I_2 [cf.(3.43)] using two independent sets of N random points, and then assert the result (3.44) for I. An alternate way is to make three simultaneous "parallel" Monte Carlo calculations of I_1, I_2 and I [cf. (3.43) and (3.45)] using a single set of N random points. In practice, the second method is usually more efficient. The fact that it requires generating only half as many points as the first method usually more than compensates for any excess of σ over $\sqrt{\sigma_1^2+\sigma_2^2}$; furthermore, σ sometimes turns out to be considerably less than $\sqrt{\sigma_1^2+\sigma_2^2}$.

The relation of σ to σ_1 and σ_2 is interesting, and deserves further discussion. Of course, in any actual "parallel" Monte Carlo calculation, we would simply approximate σ_1, σ_2 and σ by

$$\sigma_k^2 \simeq \langle f_k^2:P_\Omega\rangle_N - \langle f_k:P_\Omega\rangle_N^2, \quad k = 1 \text{ and } 2$$

and

$$\sigma^2 \simeq \langle f^2 : P_\Omega \rangle_N - \langle f : P_\Omega \rangle_N^2$$

and let these values speak for themselves. However, it is instructive to examine under what conditions and by how much σ^2 will be greater than or less than $\sigma_1^2 + \sigma_2^2$. In Appendix D we prove that [cf. (D.5)]

$$\sigma^2 = \sigma_1^2 + \sigma_2^2 + 2\mathrm{cov}(f_1, f_2 : P_\Omega) \tag{3.46}$$

where the <u>covariance</u> of $f_1(\vec{x})$ and $f_2(\vec{x})$ with respect to $P_\Omega(x)$ is defined by

$$\mathrm{cov}(f_1, f_1 : P_\Omega) \equiv \langle f_1 f_2 : P_\Omega \rangle - \langle f_1 : P_\Omega \rangle \langle f_2 : P_\Omega \rangle \tag{3.47}$$

Evidently, then, σ^2 will be greater or less than $\sigma_1^2 + \sigma_2^2$ according to whether $\mathrm{cov}(f_1, f_2 : P_\Omega)$ is positive or negative. We also show in Appendix D that $\mathrm{cov}(f_1, f_2 : P_\Omega)$ is bounded according to [cf. (D.9)]

$$-\sigma_1 \sigma_2 \leqslant \mathrm{cov}(f_1, f_2 : P_\Omega) \leqslant +\sigma_1 \sigma_2 \tag{3.48}$$

Inserting this into (3.46) yields the inequality

$$|\sigma_1 - \sigma_2| \leqslant \sigma \leqslant \sigma_1 + \sigma_2 \tag{3.49}$$

We may interpret these results as follows: If the functions $f_1(\vec{x})$ and $f_2(\vec{x})$ are such that $\mathrm{cov}(f_1, f_2 : P_\Omega)$ assumes its maximum possible value of $+\sigma_1\sigma_2$, then $f_1(\vec{x})$ and $f_2(\vec{x})$ are said to be "maximally positively correlated", and the rms deviation σ of the set $\{f_1(\vec{x}_i) + f_2(\vec{x}_i)\}$ is equal to $\sigma_1 + \sigma_2$. If the functions $f_1(\vec{x})$ and $f_2(\vec{x})$ are such that $\mathrm{cov}(f_1, f_2 : P_\Omega) = 0$, then the rms deviation σ of the set $\{f_1(\vec{x}_i)+f_2(\vec{x}_i)\}$ is equal to $\sqrt{\sigma_1^2+\sigma_2^2}$; this is the same as one would get by forming the sums of *independently* chosen numbers from the two sets $\{f_1(\vec{x}_i)\}$ and $\{f_2(\vec{x}_i)\}$. Finally, if the functions $f_1(\vec{x})$ and $f_2(\vec{x})$ are such that $\text{cov}(f_1,f_2{:}P_\Omega)$ assumes its minimum possible value of $-\sigma_1\sigma_2$, then $f_1(\vec{x})$ and $f_2(\vec{x})$ are said to be "maximally negatively correlated", and the rms deviation σ of the set $\{f_1(\vec{x}_i)+f_2(\vec{x}_i)\}$ is equal to $|\sigma_1-\sigma_2|$.

In the present instance, it is clearly advantageous for $f_1(\vec{x})$ and $f_2(\vec{x})$ to be *negatively* correlated; for then the uncertainty in the estimate of I obtained using a *single* set of N points via (3.45) would actually be *less* than the uncertainty obtained using *two* sets of N points via (3.44). Generally speaking, $f_1(\vec{x})$ and $f_2(\vec{x})$ will be negatively correlated if the minima of $f_1(\vec{x})$ tend to occur in regions where $f_2(\vec{x})$ has its maxima, and vice-versa.

3-5. Monte Carlo versus Other Numerical Integration Methods

At this point it seems appropriate to make a few brief comments on the relative advantages and disadvantages of the Monte Carlo method as compared with the more conventional numerical integration methods.

If the integral is one-dimensional and the integrand is fairly smooth, then the standard quadrature methods are far superior to Monte Carlo. For example, the relatively crude Trapezoid Rule will have in this case an associated uncertainty which decreases like $1/N^2$ as the number N of (evenly spaced) sampling points increases; by contrast, the uncertainty associated with the Monte Carlo method decreases only like $1/N^{1/2}$ as the number N of (random) sampling points increases. However, for integrals of high dimensionality the situation can be otherwise. For example, in d dimensions the uncertainty inherent in the Trapezoid Rule decreases only as $1/N^{2/d}$ whereas the uncertainty in the Monte Carlo method *remains* proportional to $1/N^{1/2}$. Moreover the extensions of the conventional methods to higher dimensions are usually quite complicated, whereas the Monte Carlo procedure is rather insensitive to the dimensionality of the integral (particularly if the integration region Ω is box-like). It should also be pointed out that the conventional methods and their attendant estimates of the uncertainty usually require the integrand to be a fairly "smooth" function, whereas the Monte Carlo method easily accommodates finite step discontinuities of the type frequently occurring in integrals of physical interest. In light of these considerations, it is seen that *the Monte Carlo method can be a very sensible choice for complicated, multi-dimensional integrals*. A rough rule-of-thumb is that, for integrals which cannot be conveniently reduced analytically below a dimensionality of 4 or 5, the Monte Carlo method should be given very serious consideration.

A particularly attractive feature about the Monte Carlo method is its operational simplicity, and the relative ease with which it produces not only the estimate of the integral but also the uncertainty in this estimate. For this reason, it sometimes makes good sense to use the Monte Carlo method even when it is *not* the most "efficient" method (in the sense of producing an answer of given accuracy using a minimum of *computer* time), for one can often obtain a sufficiently accurate answer at an acceptable computer cost using Monte Carlo much faster than one can implement a more efficient but more complicated standard numerical method. And even if it is desired to have the normally greater accuracy of the conventional methods when the dimensionality is less than 5, a simple Monte Carlo calculation can often provide a reassuring independent check against gross errors.

Finally it should be noted that, from the standpoint of the computer, the Monte Carlo method makes very minimal demands upon storage capacity and input/output devices. In particular, one does not have to store lots of points $\vec{x}$ and/or their associated f-values [see Fig. 9]. Thus, when executing a Monte Carlo program, the computer will normally be "compute bound", or limited only by the speed with which it can perform standard arithmetical and logical operations. From a strictly financial point of view (which may well be the most sensible measure of "efficiency") one should therefore pay attention to whether one's computer charges are calculated on the basis of "cpu time" (i.e., the time actually spent by the central processing unit in carrying out the required arithmetical and logical operations), or on the basis of "core time" (i.e., the cpu time weighted by the memory storage used and the number of times input/output devices are accessed). For a typical compute-bound Monte Carlo calculation, core-time charges can range from 2/3 to only 1/10 of the cpu-time charges. Until recently most computer centers charged on the basis of cpu time; however, for the newer computers which run in a time-sharing mode, core time has been shown to provide a more realistic and equitable basis for charging users. As a result, most large computer centers now charge on the basis of core time, a fact which makes the Monte Carlo method today even more attractive.

Despite the foregoing advantages of the Monte Carlo method in certain circumstances, it nevertheless frequently happens that one plays the Monte Carlo game through according to the rules, but finds that one's answer has an uncertainty which is simply too large. Since increasing N (and therefore the computer running time, and therefore the cost) by a factor of k decreases the uncertainty by a factor of only $1/\sqrt{k}$, one is tempted in such cases to discard the Monte Carlo method as unsuitable. While this may indeed be the appropriate course of action, one should not take this step without giving some consideration to the variance reducing techniques which we shall outline in Chapter 4. Essentially, these techniques try to decrease the numerator of the uncertainty in (3.21) without significantly increasing the required computer time. It is probably fair to say that the relatively recent recognition and use of these variance reducing techniques has, more than anything else, served to elevate Monte Carlo to the level of a "respectable colleague" of the conventional numerical quadrature methods.

Chapter 4

VARIANCE REDUCING TECHNIQUES

4-1. General Considerations

We have seen that the straightforward Monte Carlo method of estimating the integral

$$I \equiv \int_{\Omega} f(\vec{x})d\vec{x} \tag{4.1}$$

consists of first picking N points $\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_N$ from the set of random points $\{\vec{x}_i\}$ distributed uniformly over Ω—i.e., according to the density function $P_\Omega(\vec{x})$ in (3.4)—and then putting

$$I \simeq \tilde{I} \pm \Delta \tag{4.2}$$

where the estimate $\tilde{I}$ is

$$\tilde{I} = |\Omega|\frac{1}{N}\sum_{i=1}^{N} f(\vec{x}_i) \tag{4.3}$$

and the uncertainty Δ is

$$\Delta = |\Omega| \frac{\sqrt{\mathrm{var}(f:P_\Omega)}}{\sqrt{N}} \tag{4.4}$$

In (4.4), $\mathrm{var}(f:P_\Omega)$ is just the variance σ^2 of the set of random numbers $\{f(\vec{x}_i)\}$[see (3.16)],

$$\mathrm{var}(f:P_\Omega) \equiv \langle f^2:P_\Omega\rangle - \langle f:P_\Omega\rangle^2 \tag{4.5}$$

which in actual computations is always approximated according to [see (3.17)]

$$\mathrm{var}(f:P_\Omega) \simeq \frac{1}{N}\sum_{i=1}^{N} f^2(\vec{x}_i) - \left(\frac{1}{N}\sum_{i=1}^{N} f(\vec{x}_i)\right)^2 \qquad (4.6)$$

It is clear from (4.4) that the uncertainty Δ can in principle be made as small as desired by taking N large enough. However, since the time required to perform the calculations is roughly proportional to N, then in practice the size of N is limited by the amount of computer time available; for example, in order to halve the Monte Carlo uncertainty obtained in a one hour computer run, we would have to perform a four hour computer run. Clearly, then, beyond a certain point it is simply not feasible to decrease the uncertainty by increasing N. It would seem that the only alternative to increasing N would be to somehow modify f and/or P_Ω in such a way that the value of I is essentially left unchanged, while the quantity $\mathrm{var}(f:P_\Omega)$ gets replaced by something smaller. Several general procedures have been devised for accomplishing this, and in the present chapter we shall give a brief discussion of four of these so-called "variance reducing" techniques. Whether or not any of these techniques can be profitably utilized in any given instance will depend very strongly upon the specific form of the integrand $f(\vec{x})$ and the integrating region Ω, as well as upon the resourcefulness of the person doing the calculation. For this reason we shall not be able to develop specific recipes for blindly applying these variance reducing techniques; all we can do is outline their basic strategies.

To get a general idea of just what is involved in "reducing the variance", it may be helpful to recall the discussion of Section 3-2. There we denoted by $\tilde{P}(y)$ the density function of the set of random numbers $\{y_i\} \equiv \{f(\vec{x}_i)\}$. In principle, the function $\tilde{P}(y)$ is uniquely determined by the two functions $f(\vec{x})$ and $P_\Omega(\vec{x})$ [see Fig. 8], but in practice it is never possible to calculate $\tilde{P}(y)$ analytically. Nevertheless, it is precisely the "center of gravity" of the curve $\tilde{P}(y)$-versus-y, namely $\langle y:\tilde{P}\rangle = \langle f:P_\Omega\rangle$, which when multiplied by $|\Omega|$ gives the sought for value of the integral I. Now, essentially what we do in a Monte Carlo calculation is to approximate the curve $\tilde{P}(y)$-versus-y by a frequency histogram of N randomly chosen numbers from the set $\{y_i\}=\{f(\vec{x}_i)\}$. Provided N is sufficiently large, the center $\langle y:\tilde{P}\rangle_N = \langle f:P_\Omega\rangle_N$ of the frequency histogram will approximate the center $\langle y:\tilde{P}\rangle = \langle f:P_\Omega\rangle$ of the $\tilde{P}(y)$ curve to within a ± uncertainty equal to the width $\sqrt{\mathrm{var}(y:\tilde{P})} = \sqrt{\mathrm{var}(f:P_\Omega)}$ of the $\tilde{P}(y)$ curve divided by $\sqrt{N}$. In an actual calculation, of course, we also approximate the width of the $\tilde{P}(y)$ curve by the width of the frequency histogram, which is given by the square root of the quantity on the right side of (4.6). It is sometimes helpful to actually plot a frequency histogram of the $y_i=f(\vec{x}_i)$ values in the course of carrying out a Monte Carlo calculation[†], since such a histogram graphically illustrates just what one is up against in obtaining an accurate Monte Carlo estimate of the integral at hand: The broader

[†] Such a histogram should be built up continuously as each new $y_i \equiv f(\vec{x}_i)$ value is obtained, rather than all at once at the end of the calculation. The idea is to avoid having to store all the y_i-values in computer memory.

this frequency histogram, the more sensitive its center will be to random fluctuations arising from the finiteness of the number N of y_i-values sampled, and hence the more uncertainty will be associated with the fundamental approximation $\langle y:\tilde{P}\rangle \simeq \langle y:\tilde{P}\rangle_N$.

Of course, a frequency histogram of the $f(\vec{x}_i)$-values should not be confused with a plot of $f(\vec{x})$-versus-$\vec{x}$. Indeed, the shapes of these two curves are sort of inversely related to each other: If $f(\vec{x})$ is relatively constant over Ω, then the frequency histogram of the $f(\vec{x}_i)$-values (or equivalently, the curve $\tilde{P}(y)$-versus-y) will consist primarily of a single, narrow peak, implying that $\mathrm{var}(f:P_\Omega)$ will be relatively small. On the other hand, if $f(\vec{x})$ is peaked and consequently assumes a wide range of values in Ω, then the frequency histogram of the $f(\vec{x}_i)$-values (or equivalently, the curve $\tilde{P}(y)$-versus-y) will be broadly spread out, implying that $\mathrm{var}(f:P_\Omega)$ will be relatively large.

Taking all these things into consideration, it is clear that any variance reducing technique must aim at modifying things in such a way that the value of the integral is unchanged, but the integrand is rendered flatter or more nearly constant over the integrating region. The consequent narrowing and sharpening of the density function $\tilde{P}(y)$ of the set of integrand values $\{y_i\}$ will make its true center easier to locate by a finite sampling procedure, thereby reducing the Monte Carlo uncertainty Δ. Indeed, if we could somehow arrange to wind up with an integrand which is perfectly constant, then the density function of the integrand values would be a single spike at that constant value with zero width, and our Monte Carlo estimate would be exact. Of course, in such a case the integral could be performed analytically, and Monte Carlo would not be needed. But the important thing here is that *the closer we can get to the ideal situation of a constant integrand the more accurate our Monte Carlo calculation is going to be*. This, in essence, is the guiding philosophy of all variance reducing techniques.

In Sections 4-2 through 4-5 we shall sketch four different strategies for reducing the variance in a Monte Carlo integration. These strategies are called *control variates*, *antithetic variates*, *stratified sampling*, and *importance sampling*. We shall see that, despite these somewhat esoteric names, the underlying principle of each method is really quite straightforward. This writer's practical experience has been mainly with the last technique (importance sampling), and in Section 4-6 we shall describe a crude but often effective procedure for applying that technique in a rather routine way.

4-2. Control Variates

Suppose we can find a function $f_0(\vec{x})$ whose integral over Ω is known exactly:

$$I_0 = \int_\Omega f_0(\vec{x})d\vec{x} \tag{4.7}$$

Then the given integral I in (4.1) can be written

$$I = I_0 + \int_\Omega [f(\vec{x})-f_0(\vec{x})]d\vec{x} \tag{4.8}$$

Now, *if* $f_0(\vec{x})$ *has a sufficiently strong correlation with* $f(\vec{x})$, so that $f_0(\vec{x})$ tends to be large where $f(\vec{x})$ is large and small where $f(\vec{x})$ is small, then the function $[f(\vec{x})-f_0(\vec{x})]$ will be more nearly constant over Ω than $f(\vec{x})$ alone is. In such a circumstance, the integral I can evidently be determined more accurately through (4.8) by performing a Monte Carlo integration of $[f(\vec{x})-f_0(\vec{x})]$ instead of by Monte Carlo integrating $f(\vec{x})$ itself. This strategy is known as the "control variates" method: essentially, the fluctuations in the variate $f(\vec{x})$, whose mean over Ω is *not* known, are to some extent "controlled" by the fluctuations in the variate $f_0(\vec{x})$, whose mean over Ω *is* known.

More quantitatively, the ratio of the uncertainty Δ^* in calculating with (4.8) to the uncertainty Δ in calculating with (4.1), assuming the same number of random points in Ω are used, is evidently

$$\frac{\Delta^*}{\Delta} = \sqrt{\frac{\mathrm{var}(f-f_0:P_\Omega)}{\mathrm{var}(f:P_\Omega)}} \qquad (4.9)$$

Now, according to (D.6)

$$\mathrm{var}(f-f_0:P_\Omega) = \mathrm{var}(f:P_\Omega) + \mathrm{var}(f_0:P_\Omega) - 2\mathrm{cov}(f,f_0:P_\Omega)$$

where the covariance of two functions with respect to a given set of random points is defined and discussed in Appendix D. Hence, we will evidently have $\Delta^*<\Delta$ *provided*

$$\mathrm{cov}(f,f_0:P_\Omega) > \tfrac{1}{2}\mathrm{var}(f_0:P_\Omega) \qquad (4.10)$$

This last inequality tells us precisely how strong the correlation between

$f_0(\vec{x})$ and $f(\vec{x})$ must be in order for (4.8) to yield a more accurate Monte Carlo estimate of I than (4.1). Of course, in practice it is never possible to ascertain beforehand whether or not a chosen function $f_0(\vec{x})$ satisfies this requirement, because $\mathrm{cov}(f,f_0:P_\Omega)$ is defined in terms of integrals which are generally more complicated than I itself [see (D.4)]. Therefore, in an actual calculation one would have to be content with finding an integrable function $f_0(\vec{x})$ whose maxima and minima roughly correspond to those of $f(\vec{x})$. A short test Monte Carlo run could then resolve the question of whether or not $\mathrm{var}(f-f_0:P_\Omega)$ is significantly less than $\mathrm{var}(f:P_\Omega)$ simply by directly estimating these two quantities in the usual way [see (4.6)].

We see, then, that the difficulty in applying the control variates method lies not in determining whether (4.10) holds for a given function $f_0(\vec{x})$, but rather in discovering a suitable $f_0(\vec{x})$ in the first place. On the one hand, $f_0(\vec{x})$ must be simple enough that its integral over Ω can be calculated analytically; on the other hand, $f_0(\vec{x})$ must be intricate enough to follow the major ups and downs of the presumably complicated function $f(\vec{x})$. Therefore, the practical limitations on the control variates method are essentially those imposed by one's limited knowledge of the detailed "shape" of the given function $f(\vec{x})$ over Ω, as well as one's limited ability to find or construct exactly integrable functions $f_0(\vec{x})$ of a similar shape.

4-3. Antithetic Variates

Suppose we can find a function $f_0(\vec{x})$ whose integral over Ω is known to be equal to the integral of the given function $f(\vec{x})$ over Ω (even though, of course, the numerical value of that common integral is unknown):

$$\int_\Omega f_0(\vec{x})d\vec{x} = \int_\Omega f(\vec{x})d\vec{x} = I \tag{4.11}$$

Then the given integral I in (4.1) can be written

$$I = \int_\Omega \frac{1}{2}[f(\vec{x})+f_0(\vec{x})]d\vec{x} \tag{4.12}$$

Now, <u>if</u> $f_0(\vec{x})$ <u>has a sufficiently strong anti-correlation with</u> $f(\vec{x})$, so that $f_0(\vec{x})$ tends to be large where $f(\vec{x})$ is small and small where $f(\vec{x})$ is large, then the function $\frac{1}{2}[f(\vec{x})+f_0(\vec{x})]$ will be more nearly constant over Ω than $f(\vec{x})$ is. In such a circumstance, the integral I can evidently be determined more accurately through (4.12) by performing a Monte Carlo integration of $\frac{1}{2}[f(\vec{x})+f_0(\vec{x})]$ instead of by Monte Carlo integrating $f(\vec{x})$ itself. This strategy is known as the "antithetic variates" method: essentially, the fluctuations in the variate $f(\vec{x})$ tend to be cancelled by the opposing fluctuations in the variate $f_0(\vec{x})$, with the result that the fluctuations in the variate $\frac{1}{2}[f(\vec{x})+f_0(\vec{x})]$ are smaller than either.

More quantitatively, the ratio of the uncertaintly Δ^* in calculating with (4.12) to the uncertainty Δ in calculating with (4.1), assuming the same number of random points in Ω are used, is evidently

$$\frac{\Delta^*}{\Delta} = \sqrt{\frac{\mathrm{var}(\frac{1}{2}[f+f_0]:P_\Omega)}{\mathrm{var}(f:P_\Omega)}} \tag{4.13}$$

Now, according to (D.6)

$$\begin{aligned}\mathrm{var}(\tfrac{1}{2}[f+f_0]:P_\Omega) &= \tfrac{1}{4}\mathrm{var}(f:P_\Omega) + \tfrac{1}{4}\mathrm{var}(f_0:P_\Omega) + \tfrac{1}{2}\mathrm{cov}(f,f_0:P_\Omega) \\ &= \mathrm{var}(f:P_\Omega) - \tfrac{3}{4}\mathrm{var}(f:P_\Omega) + \tfrac{1}{4}\mathrm{var}(f_0:P_\Omega) \\ &\quad + \tfrac{1}{2}\mathrm{cov}(f,f_0:P_\Omega)\end{aligned}$$

Hence, we will evidently have $\Delta^* < \Delta$ provided

$$\mathrm{cov}(f,f_0:P_\Omega) < \tfrac{3}{2}\mathrm{var}(f:P_\Omega) - \tfrac{1}{2}\mathrm{var}(f_0:P_\Omega) \tag{4.14}$$

This last inequality tells us precisely how strong the anticorrelation between $f_0(\vec{x})$ and $f(\vec{x})$ must be in order for (4.12) to yield a more accurate Monte Carlo estimate of I than (4.1). Of course, in practice, it is never possible to ascertain beforehand whether or not a chosen function $f_0(\vec{x})$ satisfies this requirement, because $\mathrm{cov}(f,f_0:P_\Omega)$ is defined in terms of integrals which are generally more complicated than I itself. Therefore, in an actual calculation one would have to be content with finding some function $f_0(\vec{x})$ whose integral over Ω is known to equal I and whose maxima and minima roughly correspond to the respective minima and maxima of $f(\vec{x})$. A short test Monte Carlo run could then resolve the question of whether or not $\mathrm{var}(\frac{1}{2}[f+f_0]:P_\Omega)$ is significantly less than $\mathrm{var}(f:P_\Omega)$ simply by directly estimating these two quantities in the usual way [see (4.6)].

Clearly, the difficulty in applying the antithetic variates method lies in finding a suitable function $f_0(\vec{x})$, just as in the control variates method. On the surface it might seem that it would be exceedingly difficult to find a function which, on the one hand, has the same integral over Ω as the given function $f(\vec{x})$, while on the other hand is strongly anticorrelated with $f(\vec{x})$. Usually the most feasible way to proceed is to define $f_0(\vec{x})$ in terms of $f(\vec{x})$ itself. As a simple illustration of how this can be done, consider the Monte Carlo evaluation of the one-dimensional integral

$$I = \int_a^b f(x)dx \tag{4.15a}$$

Suppose we define $f_0(x)$ by

$$f_0(x) \equiv f(a+b-x) \tag{4.15b}$$

That this function $f_0(x)$ satisfies the fundamental requirement that its integral from a to b equals I is easily proved by changing integration variables according to $x \rightarrow x' = a+b-x$. Thus we have as in (4.12)

$$I = \int_a^b \frac{1}{2}[f(x) + f_0(x)]dx = \int_a^b \frac{1}{2}[f(x) + f(a+b-x)]dx \tag{4.15c}$$

Now, _if_ it happens that $f(x)$ is monotonically increasing (decreasing) in $a<x<b$, then $f_0(x)$ will be monotonically decreasing (increasing) in $a<x<b$; as a consequence, the integrand in (4.15c) will be more nearly constant over $a<x<b$ than $f(x)$, and will thus have a smaller variance. Indeed, if $f(x)$ were the simple _linear_ function $Ax+B$, then the integrand in (4.15c) would be a _constant_, and the variance would be _zero_.

In less trivial multidimensional applications, one can try to construct a suitable function $f_0(\vec{x})$ in terms of $f(\vec{x})$ in an analogous way.

Specifically, one puts

$$f_0(\vec{x}) \equiv f(\vec{x}') \tag{4.16a}$$

where $\vec{x}\to\vec{x}'$ is a transformation which satisfies the two conditions

$$\left.\begin{array}{l} \bullet\ \vec{x}\to\vec{x}' \text{ maps } \Omega \text{ onto itself} \\ \bullet\ \text{the Jacobian } \dfrac{\partial\vec{x}}{\partial\vec{x}'} = 1 \end{array}\right\} \tag{4.16b}$$

These two conditions insure that the integral of $f_0(\vec{x})$ over Ω is equal to the integral of $f(\vec{x})$ over Ω, since

$$\int_\Omega f_0(\vec{x})d\vec{x} \equiv \int_\Omega f(\vec{x}')d\vec{x} = \int_\Omega f(\vec{x}')\left|\frac{\partial\vec{x}}{\partial\vec{x}'}\right|d\vec{x}' = \int_\Omega f(\vec{x}')d\vec{x}'$$

The remaining details of the transformation $\vec{x}\to\vec{x}'$ [that is, all properties not specified by (4.16b)] are then chosen in such a way that the transformation tends to carry points for which $f(\vec{x})$ is large into points for which $f(\vec{x})$ is small, and vice-versa; this will result in $f_0(\vec{x})$ in (4.16a) being "anticorrelated" with $f(\vec{x})$. Clearly, one needs to know a good deal about the behavior of $f(\vec{x})$ in Ω in order to devise such a transformation which will anticorrelate $f_0(\vec{x})$ and $f(\vec{x})$ to a sufficient degree that $\text{var}(\frac{1}{2}(f_0+f):P_\Omega)$ will be significantly less than $\text{var}(f:P_\Omega)$.

4-4. Stratified Sampling

Let the given integrating region Ω be partitioned into n subregions $\Omega_1,\Omega_2,\ldots,\Omega_n$. Then the integral I in (4.1) can be written

$$I = \sum_{j=1}^{n} I_j \qquad (4.17)$$

where I_j is the integral of $f(\vec{x})$ over the subregion Ω_j:

$$I_j \equiv \int_{\Omega_j} f(\vec{x})d\vec{x}, \quad j=1,2,\ldots,n \qquad (4.18)$$

Suppose now that we Monte Carlo integrate each integral I_j separately; that is, we pick N_j random points $\vec{x}_1,\vec{x}_2,\ldots,\vec{x}_{N_j}$ from the uniform distribution inside Ω_j, and we put [cf. (4.2)-(4.4)]

$$I_j \simeq \tilde{I}_j + \Delta_j \qquad (4.19)$$

where

$$\tilde{I}_j = |\Omega_j|\frac{1}{N_j} \sum_{i=1}^{N_j} f(\vec{x}_i) \qquad (4.20)$$

and

$$\Delta_j = |\Omega_j|\frac{\sqrt{\mathrm{var}(f:P_{\Omega_j})}}{\sqrt{N_j}} \qquad (4.21)$$

Since these Monte Carlo estimates of the n I_j integrals are statistically independent of each other, then we can obtain a Monte Carlo estimate of the sum of the I_j integrals, namely I, by applying the prescription in equations (3.37)-(3.39). Specifically, we may assert that the integral I in (4.1) is given approximately by the estimate

$$\tilde{I} = \tilde{I}_1 + \tilde{I}_2 + \ldots + \tilde{I}_n \qquad (4.22a)$$

and further that the ± Gaussian uncertainty associated with this estimate is

$$\Delta^* = \sqrt{\Delta_1^2 + \Delta_2^2 + \ldots + \Delta_n^2} \tag{4.22b}$$

The foregoing method of estimating the integral I is called the "stratified sampling" method (for reasons to become apparent later). It is obviously a legitimate way of proceeding, but the extra effort involved is clearly pointless unless the uncertainty Δ^* in (4.22b) is significantly less than the uncertainty Δ in (4.4), given that the total number of points used in the two procedures are the same. The question of interest, then, is as follows: Given that

$$\sum_{j=1}^{n} N_j = N \tag{4.23}$$

under what conditions (if any) will

$$\Delta^{*2} = \sum_{j=1}^{n} |\Omega_j|^2 \frac{\mathrm{var}(f:P_{\Omega_j})}{N_j} \tag{4.24}$$

be significantly less than

$$\Delta^2 = |\Omega|^2 \frac{\mathrm{var}(f:P_\Omega)}{N} \tag{4.25}$$

In addressing this question let us begin by considering a specific situation which, although somewhat contrived, would obviously be handled more efficiently by the stratified sampling method than by the ordinary sampling method. Thus, suppose $f(\vec{x})$ consists of a number of perfectly flat "plateaus" over Ω, in that

$$f(\vec{x}_j) = C_j \quad \text{for} \quad \vec{x}_j \varepsilon \Omega_j \quad (j=1,2,\ldots,n)$$

where $\Omega_1, \Omega_2, \ldots, \Omega_n$ are a particular set of n non-overlapping subregions whose union is Ω. In such a situation we would evidently have $\text{var}(f:P_{\Omega_j}) = 0$ for each j, so that the uncertainty Δ^* in (4.24) associated with a stratified sampling over the corresponding regions $\Omega_1, \ldots, \Omega_n$ would vanish. On the other hand, the uncertainty Δ in (4.25) associated with the regular method would not vanish unless $f(\vec{x})$ were constant everywhere inside Ω—i.e., unless the constants C_j were all equal to each other.

To illustrate this situation more concretely, the reader can easily verify that an ordinary N-point Monte Carlo calculation of the integral over $\Omega = (0,1)$ of the step function

$$f(x) = \begin{cases} 1, & \text{for} \quad 0 \leqslant x \leqslant 1/2 \\ 2, & \text{for} \quad 1/2 < x \leqslant 1 \end{cases}$$

will have an associated uncertainty of

$$\Delta = |1| \frac{\sqrt{\text{var}(f:P_\Omega)}}{\sqrt{N}} = \frac{1/2}{\sqrt{N}}$$

However, if we partition the integrating region $\Omega = (0,1)$ into the two subregions $\Omega_1 = (0,1/2)$ and $\Omega_2 = (1/2,1)$, then since f(x) is constant inside each subregion we will have $\text{var}(f:P_{\Omega_1}) = \text{var}(f:P_{\Omega_2}) = 0$; therefore, the uncertainty Δ^* associated with a stratified sampling procedure over these two subregions will vanish.

Addressing the problem more generally, we show in Appendix E that, for any given function $f(\vec{x})$ and any partitioning of the given region Ω into subregions $\Omega_1,\Omega_2,\ldots,\Omega_n$, we have [cf. (E.4)]

$$\mathrm{var}(f:P_\Omega) = \sum_j \alpha_j \mathrm{var}(f:P_{\Omega_j}) + \sum_{j<k}\sum \alpha_j\alpha_k(\langle f:P_{\Omega_j}\rangle - \langle f:P_{\Omega_k}\rangle)^2 \qquad (4.26a)$$

where

$$\alpha_j \equiv |\Omega_j|/|\Omega| \qquad (4.26b)$$

The two-term structure of (4.26a) reveals that the variance of $f(\vec{x})$ over Ω can be regarded as coming from two sources: (i) the variations in $f(\vec{x})$ within each of the various subregions [the first term in (4.26a)]; and (ii) the variations in $f(\vec{x})$ among the various subregions [the second term in (4.26a)]. Now, it is clear from equations (4.25) and (4.24) that, whereas both of these sources of variation contribute to Δ, only the first source contributes to Δ^*. Therefore, if we can devise a partitioning of Ω which minimizes (i)—or equivalently maximizes (ii)—then we may expect that Δ^* will be significantly less than Δ. This in essence is the guiding philosophy of the stratified sampling technique: If the integrand $f(\vec{x})$ has several fairly level plateaus or "strata", then independent samplings over the regions $\Omega_1,\Omega_2,\ldots,\Omega_n$ under each strata will result in a more accurate estimate of the integral than a sampling over the entire region Ω which ignores the strata.

Apart from the problem of determining an effective partitioning scheme for the region Ω — and this is really the key problem in applying the stratified sampling procedure — we also have the problem of deciding how to apportion the N sampling points among the various subregions Ω_j. One simple and intuitively plausible way of doing this would be to make N_j proportional to the size of Ω_j. Thus, we would put

$$N_j = \frac{N}{|\Omega|} |\Omega_j| = N\alpha_j \tag{4.27a}$$

This of course is roughly how the points would apportion themselves in an ordinary "unstratified" sampling procedure. If we insert (4.27a) into the expression for Δ^* in (4.24) and then divide by the expression for Δ in (4.25), we find that this way of apportioning the points produces the following ratio of Δ^* to Δ:

$$\frac{\Delta^*}{\Delta} = \sqrt{\frac{\sum_j \alpha_j \mathrm{var}(f:P_{\Omega_j})}{\mathrm{var}(f:P_{\Omega})}} \tag{4.27b}$$

Now, since the second term in (4.26a) cannot be negative, we see that the numerator in (4.27b) cannot exceed the denominator [cf. (E.5)]. Thus we conclude that, if the numbers N_j are chosen according to the simple prescription (4.27a), then <u>regardless</u> of how astutelv the partitioning is chosen we shall have $\Delta^* \leqslant \Delta$. However, it is also clear from (4.26a) that the more care one takes to exploit any "plateau-like" behavior of $f(\vec{x})$ in choosing the partitioning, the smaller the ratio Δ^*/Δ is going to be.

Actually, the apportioning of points to each subregion strictly according to the size of the subregion is <u>not</u> the optimum procedure, even though as we have seen it can never result in making $\Delta^* > \Delta$. In Appendix F we show that the best procedure is to make N_j proportional to the size of Ω_j times the rms variation of $f(\vec{x})$ over Ω_j [cf. (F.9)]:

$$N_j = K|\Omega_j|\sqrt{\mathrm{var}(f:P_{\Omega_j})} \qquad (4.28a)$$

Here the constant K is to be determined by the requirement $\sum_j N_j = N$. Of course, we would normally not have any a priori knowledge of the quantities $\mathrm{var}(f:P_{\Omega_j})$. In practice, therefore, a sensible procedure to follow would be to apportion the points according to (4.27a) in a short preliminary calculation, and then adjust the apportiorment more along the lines of (4.28a) on the basis of the estimates for $\mathrm{var}(f:P_{\Omega_j})$ obtained in the preliminary calculation. By combining (4.28a) with (4.24) and (4.25), it is a simple matter to show that this optimum apportionment of sampling points leads to [cf. (F.10)]

$$\frac{\Delta^*}{\Delta} = \frac{\sum_j \alpha_j \sqrt{\mathrm{var}(f:P_{\Omega_j})}}{\sqrt{\mathrm{var}(f:P_\Omega)}} \qquad (4.28b)$$

instead of (4.27b).

Again, we should emphasize that the major problem presented by the stratified sampling method is to discover a sensible way of partitioning Ω into subregions $\Omega_1, \Omega_2, \ldots$. Generally speaking, the method is worthwhile only if the "curve" $f(\vec{x})$-versus-$\vec{x}$ exhibits a quasi-plateaued appearance. One must then be able to associate with each plateau a subregion Ω_j whose shape is sufficiently simple that: (i) its volume

$|\Omega_j|$ can be calculated exactly, and (ii) random points can easily be generated uniformly inside Ω_j. As with the control variates and the antithetic variates methods, a successful application of the stratified sampling method clearly requires a fairly detailed knowledge of the behavior of $f(\vec{x})$ in Ω.

4-5. Importance Sampling (Theory)

Let $P(\vec{x})$ by any probability density function which is normalized on, and non-zero in, the given integrating region Ω:

$$\int_\Omega P(\vec{x})d\vec{x} = 1 \tag{4.29a}$$

$$P(\vec{x}) > 0 \quad \text{for all } \vec{x}\varepsilon\Omega \tag{4.29b}$$

The latter property allows us to multiply and divide the integrand in (4.1) by $P(\vec{x})$ to obtain the following equivalent expression for the integral I:

$$I = \int_\Omega [f(\vec{x})/P(\vec{x})]P(\vec{x})d\vec{x} \tag{4.30}$$

With (3.3) we can write this as

$$I = \langle f/P:P\rangle \tag{4.31}$$

which says that the integral (4.1) can be regarded as the average of $f(\vec{x})/P(\vec{x})$, taken with respect to the set of random points $\{\vec{x}_i\}$ distributed according to the density function $P(\vec{x})$. From a Monte Carlo point of view this implies that if $\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_N$ are N random points <u>distributed according to the density function</u> $P(\vec{x})$, then we can put

$$I \simeq \tilde{I}^* \pm \Delta^* \qquad (4.32)$$

where

$$\tilde{I}^* = \frac{1}{N}\sum_{i=1}^{N} f(\vec{x}_i)/P(\vec{x}_i) \qquad (4.33)$$

and

$$\Delta^* = \frac{\sqrt{\mathrm{var}(f/P:P)}}{\sqrt{N}} \qquad (4.34)$$

We remark again that, except for the requirements (4.29), the form of the density function $P(\vec{x})$ is quite arbitrary. One possible choice for $P(\vec{x})$ is of course the <u>uniform</u> density function $P_\Omega(\vec{x})$ in (3.4). For this choice we have $f(\vec{x})/P(\vec{x}) = |\Omega| f(\vec{x})$ everywhere inside Ω, so that (4.31) reduces to (3.5), and (4.32)-(4.34) reduce to (4.2)-(4.4), respectively. These latter equations have formed the basis for most of our discussion of the Monte Carlo method of evaluating integrals. However, the foregoing observations suggest that we need not be inextricably wedded to the uniform distribution. As we shall see below, whereas the average $\tilde{I}^*$ in (4.33) is essentially independent of the form of $P(\vec{x})$, the uncertainty Δ^* in (4.34) depends rather strongly on the form of $P(\vec{x})$. Therefore, if the uncertainty which results from the usual procedure of taking $P(\vec{x}) = P_\Omega(\vec{x})$ [i.e., Δ in (4.4)] turns out to be unacceptably large, it might be possible to use some other form for $P(\vec{x})$ and thereby reduce the uncertainty.

Let us first verify that $\tilde{I}^*$ in (4.33) is, in the limit of large N, independent of the form of $P(\vec{x})$. Essentially, this follows from the observation that $\tilde{I}^*$ in (4.33) is by definition $\langle f/P:P\rangle_N$, and

$$\lim_{N\to\infty}\langle f/P:P\rangle_N = \langle f/P:P\rangle = \int_\Omega [f(\vec{x})/P(\vec{x})]P(\vec{x})d\vec{x}$$

so

$$\lim_{N\to\infty}\langle f/P:P\rangle_N = \int_\Omega f(\vec{x})d\vec{x} \tag{4.35}$$

To appreciate this important point in the context of an actual Monte Carlo process, let us imagine changing from $P(\vec{x})$ to some new density function $P'(\vec{x})$. Suppose that, inside a given infinitesimal region $d\Omega$ of Ω, $P'(\vec{x})$ is k times as large as $P(\vec{x})$:

$$P'(\vec{x}) = kP(\vec{x}), \quad \text{for } \vec{x}\varepsilon d\Omega$$

This implies that, in the limit of large N, we shall sample k times as many random points $\vec{x}_i$ inside $d\Omega$ with P' as with P. However, each such point sampled with P' will contribute to the sum in (4.33) the amount

$$f(\vec{x})/P'(\vec{x}) = f(\vec{x})/[kP(\vec{x})] = \frac{1}{k}[f(\vec{x})/P(\vec{x})], \quad \text{for } \vec{x}\varepsilon d\Omega$$

which is precisely 1/k times the contribution of each point sampled inside $d\Omega$ with P. Clearly, if we sample k times as many points inside $d\Omega$ while weighting the contribution of each such point by a factor 1/k, then the net contribution of $d\Omega$ to the sum in (4.33) will be unchanged. Applying this argument to every infinitesimal subregion $d\Omega$ of Ω [allowing the value of k to vary among these infinitesimal subregions in accordance with the behavior of the functions $P(\vec{x})$ and $P'(\vec{x})$], we thus see that $\tilde{I}^*$ in (4.33) is indeed insensitive to the form of $P(\vec{x})$.

Such is not the case, however, with Δ^* in (4.34). From a purely formal point of view we have

$$\begin{aligned}\mathrm{var}(f/P:P) &= \langle (f/P)^2 : P \rangle - \langle f/P : P \rangle^2 \\ &= \int_\Omega [f(\vec{x})/P(\vec{x})]^2 P(\vec{x})d\vec{x} - \left[\int_\Omega [f(\vec{x})/P(\vec{x})]P(\vec{x})d\vec{x}\right]^2\end{aligned}$$

Thus

$$\mathrm{var}(f/P:P) = \int_\Omega f^2(\vec{x})P^{-1}(\vec{x})d\vec{x} - I^2 \qquad (4.36)$$

which shows that var(f/P:P), and hence Δ^* in (4.34), indeed depends upon the functional form of $P(\vec{x})$. The question now is, how can we choose $P(\vec{x})$ to minimize the uncertainty?

Qualitatively, the answer to this question is rather obvious: To minimize the variance of $f(\vec{x})/P(\vec{x})$ with respect to <u>any</u> set of random points, we must choose $P(\vec{x})$ so as to make $f(\vec{x})/P(\vec{x})$ as constant as possible. The more nearly constant $f(\vec{x})/P(\vec{x})$ is, the smaller the variations will be among the individual terms in (4.33), and the less will be the uncertainty associated with their average $\tilde{I}^*$.

More quantitatively, we prove in Appendix G that the density function $P_{min}(\vec{x})$ which minimizes var(f/P:P), and hence Δ^*, is[†]

[†] The results (4.37) and (4.38) were apparently first derived by H. Kahn (Ref. 9).

$$P_{\min}(\vec{x}) = |f(\vec{x})| \Big/ \int_{\Omega} |f(\vec{x}')| d\vec{x}' \tag{4.37}$$

In other words, the optimum choice for $P(\vec{x})$ is, apart from a normalization constant, just the function $|f(\vec{x})|$. Inserting this optimum density function into (4.36), we find that the smallest possible value of Δ^* is [cf. (G.18b)]

$$\Delta^*_{\min} = 2N^{-1/2} \sqrt{\int_{\Omega^+} |f(\vec{x})| d\vec{x} \cdot \int_{\Omega^-} |f(\vec{x})| d\vec{x}} \tag{4.38}$$

where Ω^+ is that portion of Ω in which $f(\vec{x})$ is everywhere positive, and Ω^- is that portion of Ω in which $f(\vec{x})$ is everywhere negative. Evidently, we will have $\Delta^*_{\min}=0$ if and only if $f(\vec{x})$ never changes sign inside Ω. For example, suppose $f(\vec{x})$ were everywhere positive inside Ω. Then we could take in accordance with (4.37)

$$P(\vec{x}) = f(\vec{x}) / \int_{\Omega} f(\vec{x}') d\vec{x}' = f(\vec{x})/I \tag{4.39}$$

so that $f(\vec{x})/P(\vec{x})$ would simply be equal to the constant I everywhere inside Ω. In this case we would have $\text{var}(f/P:P)=\text{var}(I:P)=0$ [this also follows by substituting (4.39) into the right side of (4.36)], so that the uncertainty Δ^* would vanish. However, this nice state of affairs is somewhat spoiled by the following considerations: If Δ in (4.4) is indeed too large, then $f(\vec{x})$ evidently assumes a wide range of values inside Ω. In such a case, efficiency considerations would preclude generating random points $\vec{x}_i$ according to the density function in (4.39) by the <u>rejection</u> method, and we would have to use the <u>inversion</u> method.

Now, in order to use the inversion method, we must be able to calculate analytically the "normalizing constant" $1/I$ in (4.39); however, by hypothesis we cannot do this. Similar considerations apply to the general case in (4.37), and we are forced to conclude that it is in practice not feasible to choose $P(\vec{x})=P_{min}(\vec{x})$ and so achieve the minimum uncertainty in (4.38).

Nevertheless, the foregoing results do provide us with something to aim for in choosing an importance sampling density function $P(x)$: *We should try to choose* $P(\vec{x})$ *so that it follows* $|f(\vec{x})|$, *proportionately, as closely as possible*; that is, $P(\vec{x})$ should tend to be large where $|f(\vec{x})|$ is large and small where $|f(\vec{x})|$ is small. This will result in $f(\vec{x})/P(\vec{x})$ being a more constant or less varying function of $\vec{x}$ in Ω than $f(\vec{x})$ alone. As a consequence, $\mathrm{var}(f/P{:}P)$ will be smaller than $\mathrm{var}(f|\Omega|{:}P_{\Omega})$, and Δ^* in (4.34) will be smaller than Δ in (4.4).

In so choosing $P(\vec{x})$ to be large where $|f(\vec{x})|$ is large, we will evidently be "biasing" our random point generating procedure in such a way that we sample more points $\vec{x}_i$ in those regions of Ω where $|f(\vec{x})|$ is relatively large. For this reason this method of reducing the variance is called "importance sampling": we sample most intensely in the "important" regions of Ω where $f(\vec{x})$ contributes most strongly to the integral. As previously noted, this sampling bias is compensated for by dividing the value of f at each point $\vec{x}_i$ by the value of P at the same point. The result is that we get the same average as in the uniform sampling case, but since the values $f(\vec{x}_i)/P(\vec{x}_i)$ being averaged exhibit less variation than $f(\vec{x}_i)$, the uncertainty is reduced.

In importance sampling, we therefore seek a density function $P(\vec{x})$ which is such that (i) $P(\vec{x})$ follows $|f(\vec{x})|$ as closely as possible, and (ii) random points $\vec{x}_i$ can be generated according to $P(\vec{x})$ fairly easily. Ultimately of course, these two requirements are incompatible with each other: On the one hand $P(\vec{x})$ must be intricate enough to follow the major variations of the presumably complicated function $f(\vec{x})$, while on the other hand $P(\vec{x})$ must be analytically simple enough so that an efficient generating algorithm can be devised. Clearly, one must in practice strive for a reasonable compromise between these two requirements. The potential success of the method in any particular instance will therefore hinge upon one's knowledge of the behavior of the integrand, as well as upon one's ability to construct efficient algorithms for generating random numbers according to prescribed density functions.

It should be noted that it is quite possible for an importance sampling procedure to make things worse instead of better: In making $P(\vec{x})$ larger than $P_\Omega(\vec{x})$ in certain regions of Ω, we must make $P(\vec{x})$ smaller than $P_\Omega(\vec{x})$ in other regions of Ω, simply in order for $P(\vec{x})$ to satisfy the normalization condition (4.29a). In our zeal we may inadvertently make $P(\vec{x})$ so small in some region that the quantity $f^2(\vec{x})P^{-1}(\vec{x})$ in (4.36) becomes correspondingly too large, and actually increases the overall variance. In practice, therefore, one must determine a suitable function $P(\vec{x})$ by a rather cautious and tentative trial-and-error process. In the next section we shall describe in somewhat more detail one practical approach to this problem.

4-6. Importance Sampling (Application)

In this section we shall outline a specific procedure for applying the variance reducing technique described in the previous section. The procedure is rather crude, but it has the advantage of being relatively routine and easy to apply. In the author's recent work (Ref. 4) this procedure has been able to decrease Monte Carlo uncertainties by amounts equivalent to increasing the number of sampling points by anywhere from a factor of 2 to a factor of 200, depending upon the integral considered.

To apply this importance sampling procedure to the calculation of an n-dimensional integral of the form

$$I = \int_{\Omega} f(\vec{x})d\vec{x} \tag{4.40}$$

it is convenient to begin by recasting the integral as an integral over the n-dimensional unit cube,

$$I = \int_0^1 dr_1 \int_0^1 dr_2 \cdots \int_0^1 dr_n \; h(r_1, r_2, \ldots, r_n) \tag{4.41}$$

Such a recasting of the integral can always be carried out, and in fact it is essentially equivalent to "preparing" the integral for the Monte Carlo averaging process. To see this more clearly, let us consider the three-dimensional integral

$$I = \iiint_{\Omega} f(x,y,z)dxdydz \tag{4.42}$$

If one can represent the integrating region of (4.42) in the general form

$$\Omega = \{(x,y,z) \| a_1 \leqslant x \leqslant b_1, \quad a_2(x) \leqslant y \leqslant b_2(x), \ a_3(x,y) \leqslant z \leqslant b_3(x,y)\} \quad (4.43)$$

then, provided the boundary functions a_2, b_2, a_3, b_3 are simple enough, one can apply the generalized inversion method to generate random points uniformly inside Ω. Thus, as described in connection with Eqs. (2.31)-(2.34), one first "conditions" the density function $P_\Omega(x,y,z)$ in the form $P(x)P(y|x)P(z|x,y)$ and calculates the corresponding one-variable distribution functions $F(x)$, $F(y|x)$, $F(z|x,y)$. Then, where r_1, r_2, r_3 are three random numbers from a uniform distribution in the unit interval, one inverts the equations

$$\left.\begin{aligned} r_1 &= F(x) \\ r_2 &= F(y|x) \\ r_3 &= F(z|x,y) \end{aligned}\right\} \qquad (4.44)$$

to obtain a random point (x,y,z) from the uniform distribution inside Ω. [This procedure is particularly straightforward if Ω is a box, in which case all the a_i and b_i are constants; see (2.35).] Using (4.44) to generate random points uniformly inside Ω, one then proceeds to calculate I in (4.42) as $|\Omega|$ times the average of $f(x,y,z)$ over these random points. However, suppose that instead of regarding (4.44) as a set of "generating formulae", we look upon (4.44) as defining a transformation of variables. This transformation evidently carries Ω in xyz-space into the unit cube in $r_1r_2r_3$-space; moreover, it has the convenient property that its Jacobian is given simply by [cf. (2.30) and (2.32)]

$$\frac{\partial(x,y,z)}{\partial(r_1,r_2,r_3)} = |\Omega| \qquad (4.45)$$

Thus, the transformed integral

$$I = \iiint\limits_{\substack{\text{unit}\\\text{cube}}} f(x,y,z)\left|\frac{\partial(x,y,z)}{\partial(r_1,r_2,r_3)}\right| dr_1 dr_2 dr_3$$

takes the form

$$I = \int_0^1 dr_1 \int_0^1 dr_2 \int_0^1 dr_3 \; h(r_1,r_2,r_3) \tag{4.46}$$

where

$$h(r_1,r_2,r_3) = |\Omega| f(x,y,z) \tag{4.47}$$

with x,y and z now being regarded as functions of r_1,r_2 and r_3 through the inverse of Eqs. (4.44).

Alternatively, if the functions a_i and b_i in (4.43) are analytically so complicated that Eqs. (4.44) are intractable, one can try the simple "linear stretching" transformation

$$\left.\begin{aligned} r_1 &= [x-a_1]/[b_1-a_1] \\ r_2 &= [y-a_2(x)]/[b_2(x)-a_2(x)] \\ r_3 &= [z-a_3(x,y)]/[b_3(x,y)-a_3(x,y)] \end{aligned}\right\} \tag{4.48}$$

Like (4.44), this transformation also carries Ω in xyz-space into the unit cube in $r_1r_2r_3$-space; however, its Jacobian is given by

$$\frac{\partial(x,y,z)}{\partial(r_1,r_2,r_3)} = [b_1-a_1][b_2(x)-a_2(x)][b_3(x,y)-a_3(x,y)] \tag{4.49}$$

instead of (4.45). Applying this transformation to I in (4.42) we again obtain an expression of the form (4.46), except that the integrand is now given by

$$h(r_1,r_2,r_3) = [b_1-a_1][b_2(x)-a_2(x)][b_3(x,y)-a_3(x,y)]f(x,y,z) \qquad (4.50)$$

instead of (4.47), where x,y and z are functions of r_1, r_2 and r_3 through the inverse of (4.48). If one has the option of proceeding via either (4.44) or (4.48), the optimal way will normally be the one for which the function h [either (4.47) or (4.50)] is a more nearly constant function of r_1, r_2 and r_3 inside the unit cube, since this will produce a smaller uncertainty in the subsequent Monte Carlo averaging of (4.46).

If one is not able to specify Ω in the ordered form of (4.43), one can try enclosing Ω in a larger region Ω' which can be so represented; for example, Ω' might be taken to be a simple box-like region which has Ω as a subregion. For any such covering region Ω' we define

$$f'(x,y,z) \equiv \begin{cases} f(x,y,z), & \text{if } (x,y,z)\epsilon\Omega \\ \quad 0 \quad , & \text{if } (x,y,z)\notin\Omega \end{cases} \qquad (4.51a)$$

so that the integral (4.42) can be written

$$I = \iiint_{\Omega'} f'(x,y,z)dxdydz \qquad (4.51b)$$

This expression may now be reduced by either of the methods just described to the form (4.46). If this procedure too proves fruitless, then one either must find some transformation of variables $(x,y,z)\to(x',y',z')$ which carries Ω into a region Ω' for which one of the above methods can be applied [cf. (3.30)], or else one must consider abandoning the Monte Carlo approach altogether.

The foregoing observations should make it clear that all the standard methods of "preparing" the integral (4.40) for the Monte Carlo averaging process can actually be regarded as transforming the given integral to the form (4.41). The underlying reason for this is that, in the final analysis, our only input data in a Monte Carlo calculation are random numbers from the uniform distribution in the unit interval. Therefore, any n-dimensional Monte Carlo integration ultimately amounts to the calculation of the average of some function with respect to the uniform distribution of points inside the n-dimensional unit cube. From this point of view, the generating formulae [e.g., the inverses of (4.44) or (4.48)] are merely equations which help specify how the function $h(r_1,\ldots,r_n)$ is obtained from the given integrand $f(\vec{x})$.

Regarding I, then, as the integral of a function $h(r_1,\ldots,r_n)$ over the n-dimensional unit cube as in (4.41), our proposed importance sampling procedure is as follows. First we set up a computer program to calculate I by the standard Monte Carlo procedure; that is, we write a computer program to implement the steps in (3.22) and Fig. 9, but with $f(\vec{x})$ everywhere replaced by $h(r_1,\ldots,r_n)$ [via the chosen generating formulae for the components of $\vec{x}$] and with Ω everywhere replaced by the n-dimensional unit cube.

Next, we incorporate into this computer program a set of statements which keeps track of, say, the 50 highest and the 50 lowest values of the integrand $y=h(r_1,\ldots,r_n)$ encountered in the course of the calculation, along with the coordinates inside the n-dimensional unit cube where these extremal values occurred. This evidently entails setting aside $(n+1)\times100$

storage locations [50 for the highest y-values, n×50 for their associated coordinates, and likewise for the 50 lowest y-values and their coordinates], along with a block of control statements which updates these locations for each new point $(r_1,\ldots,r_n)$ generated. For example, in a three-dimensional integration one could define the variables YHI(K), R1HI(K), R2HI(K), R3HI(K) for K ranging from 1 to 50, with the understanding that YHI(K) always contains the Kth highest y-value found, and (R1HI(K), R2HI(K), R3HI(K)) the location of the point (r_1,r_2,r_3) in the unit cube where that value was found; similarly, YLO(K) would always contain the Kth lowest y-value found, and (R1LO(K), R2LO(K), R3LO(K)) the coordinates of the corresponding point. Then for each new random point (r_1,r_2,r_3) generated inside the unit cube, we not only incorporate its integrand value $y=h(r_1,r_2,r_3)$ into the cumulating sums S_1 and S_2 [cf. (3.22) and Fig. 9], but we also check to see if y is greater than YHI(50) or less than YLO(50). If, for instance, the former were found to be the case, then the current values for YHI(50), R1HI(50), R2HI(50), R3HI(50) would be discarded, and the high-values table would be shifted so as to incorporate this newest high value and its coordinates at the appropriate level.

Now we make "Preliminary Run #1", an initial computer run of the above Monte Carlo program which uses just enough random points to yield a reasonable estimate of the uncertainty Δ, as well as a reasonable sampling of the highs and lows of the integrand inside the unit cube. We next examine the coordinates of the high and low integrand values with a view to determining if these extremal values seem to be associated with a relatively narrow interval of one or more of the r_i-coordinates. To the extent that such an "important" interval on any r_i-axis can be identified, the idea is to apply an appropriate form of importance sampling on that r_i-variable <u>independently</u> of the other coordinate variables.

Suppose it is found from Preliminary Run #1 that the integrand assumes extremal values (i.e., values far from the average integrand value) whenever r_j falls inside some small subinterval (α_j,β_j) of the unit interval. We then choose some probability density function $P_j(r_j)$ which is normalized on and non-zero in the interval $0\leqslant r_j\leqslant 1$, and which has the further property that it is <u>large</u> whenever $\alpha_j<r_j<\beta_j$. In Appendix H we discuss several forms for $P_j(r_j)$ which are suitable for various intervals (α_j,β_j). Since the function $P_j(r_j)$ is non-zero in $0\leqslant r_j\leqslant 1$, we can multiply and divide the integrand in (4.41) by $P_j(r_j)$ to obtain

$$I = \int_0^1 dr_1\cdots\int_0^1 P_j(r_j)dr_j\cdots\int_0^1 dr_n\left(\frac{h(r_1,\ldots,r_j,\ldots,r_n)}{P_j(r_j)}\right) \qquad (4.52)$$

Now, in the spirit of the discussion in the first part of this section, let us make the change of variable

$$r_j \rightarrow r_j' \equiv \int_0^{r_j} P_j(r)dr \equiv F_j(r_j) \qquad (4.53)$$

where $F_j(r_j)$ is the distribution function corresponding to the density function $P_j(r_j)$. From (4.53) it is seen that as r_j ranges from 0 to 1, r_j' also ranges from 0 to 1; furthermore, it is seen that dr_j' is precisely equal to $P_j(r_j)dr_j$. Therefore, (4.52) can be written

$$I = \int_0^1 dr_1\cdots\int_0^1 dr_j'\cdots\int_0^1 dr_n\left(\frac{h(r_1,\ldots,r_j,\ldots,r_n)}{P_j(r_j)}\right) \qquad (4.54)$$

where r_j in the integrand is now to be regarded as a function of r_j' through the inverse of (4.53); that is, $r_j = F_j^{-1}(r_j')$. Comparing (4.41) with (4.54), it is clear that the Monte Carlo uncertainty associated with the latter should be smaller than the uncertainty associated with the former; for, the large value of the denominator in the integrand of (4.54) for $\alpha_j < r_j < \beta_j$ will moderate the extremal behavior of the numerator there, with the result that the variance of the integrand in (4.54) will be less than the variance of the integrand in (4.41). From a slightly different point of view, whereas in (4.41) we pick r_j randomly according to the unit uniform distribution, in (4.54) we pick r_j randomly according to the density function $P_j(r_j)$. To "correct" for this sampling bias we must introduce a factor $1/P_j(r_j)$ into the integrand, and this factor has been specifically chosen to "smooth" the integrand. This is, of course, the basic philosophy of importance sampling.[†]

[†] If the extremal coordinates list should indicate that when r_j is near 0 (or 1) the integrand assumes high extremal values, while when r_j is near 1 (or 0) the integrand assumes low extremal values, then it may be better to use a form of the antithetic variates method instead of importance sampling. To this end we introduce the function

$$h'(r_1,\ldots,r_n) \equiv \frac{1}{2}[h(r_1,\ldots,r_j,\ldots,r_n) + h(r_1,\ldots,1-r_j,\ldots,r_n)]$$

the integral of which (over the unit cube) is precisely equal to the integral of h. By taking h' as the integrand we can largely eliminate the extremal behavior near $r_j=0$ and $r_j=1$, since the two terms in brackets will tend to cancel each other whenever r_j is near 0 or 1.

We now incorporate this importance sampling procedure for r_j into our Monte Carlo computer program, so tnat the calculaticn will proceed via (4.54) instead of (4.41). This wilJ require essentially two modifications:

1° Instead of picking r_j as a random number from the uniform distribution in the unit interval, we must pick r_j' as a random number from the uniform distribution in the unit interval and then obtain r_j by inverting (4.53). [In other words, r_j is now to be picked as a random number from the distribution defined by the density function $P_j(r_j)$.]

2° Instead of taking the integrand to be $h(r_1,\ldots,r_n)$, we take the integrand to be $h(r_1,\ldots,r_n)/P_j(r_j)$.

In carrying out these two steps it is usually convenient to employ a computer subroutine which takes r_j' as input and which calculates and outputs r_j and $P_j(r_j)$. The discussion in Appendix H of several explicit importance sampling density functions is given with these subroutine requirements in mind.

It should be clear that we can carry out this single variable importance sampling procedure simultaneously for as many of the variables $r_1,r_2,\ldots,r_n$ as might seem to require it. In general we end up with an expression of the form (4.54) with all the appropriate differentials primed and with the given integrand divided by the product of all the importance sampling density functions being used. We then modify the original Monte Carlo computer program according to steps 1° and 2° above for each variable being importance sampled, and we make Preliminary Run #2, using the same number of random points as was used in Preliminary Run #1. By comparing the uncertainty obtained in the second run with that obtained in the first, we can directly assess the effect of the importance sampling. Further, by examining the coordinates of the new extremal integrand values, we can again investigate whether these values seem to be associated with a small range of any of the r_i coordinates. We may find that one or more variables not being importance sampled now seem to require it, and/or we may find that variables which are being importance sampled require some adjustment in the form of the density function being used. The latter is a particularly frequent finding, and is usually handled most expeditiously by adjusting the value of some parameter which controls the amount of density function peaking. For example, for an integrand which assumes extremal values near $r_j=0$, one might try [cf. (H.4)]

$$P_j(r_j) = \frac{\Gamma}{1 - e^{-\Gamma}} e^{-\Gamma r_j}, \quad \Gamma>0$$

the peaking of which around $r_j=0$ is roughly proportional to Γ. The idea is to take Γ large enough so that low r_j-values contribute less to the variance of the integrand; however, if Γ is taken too large, then one will find <u>high</u> r_j-values contributing extremal integrand values, thereby increasing the variance. Clearly some experimentation will be required to discover a reasonably optimal value for Γ.

Generally speaking, it will take several "preliminary runs" to settle on a good set of importance sampling density functions. One must take care not to make so many preliminary runs that one uses as much or more time than the final importance sampling scheme will save. Usually

3 to 10 preliminary runs should enable one to do about all that can be done by this method. Then, of course, one makes a final long run, or as recommended in Sec. 3-3 four final long runs, using as high a value for N as time and money will permit to obtain the final Monte Carlo estimate of the integral. To save computer storage space and execution time, one can remove the extremal values bookkeeping machinery from the program before making the final calculational run(s).

In striving to reduce the uncertainty by the foregoing importance sampling procedure, one should not expect to achieve the minimum uncertainty allowed by the general theory of importance sampling, which was discussed in the preceding section [see (4.38)]. The reason is that this simple procedure cannot effect extremal behavior caused by correlations between two or more variables. One way to see this is to observe that the most general importance sampling density function for (4.41) which can be realized by the simple method described here has the form

$$P(r_1,\ldots,r_n) = P_1(r_1)P_2(r_2)\cdots P_n(r_n) \tag{4.55a}$$

However, as we saw in the last section [see (4.37) and also Appendix G], the importance sampling density function which minimizes the Monte Carlo uncertainty for (4.41) has the form

$$P_{\min}(r_1,\ldots,r_n) = \text{const} \times |h(r_1,\ldots,r_n)| \tag{4.55b}$$

That $P(r_1,\ldots,r_n)$ in (4.55a) is not generally capable of representing $P_{\min}(r_1,\ldots,r_n)$ in (4.55b) is obvious.

Despite the fact that this simple one-variable importance sampling procedure can achieve at best only a partial reduction in the uncertainty, this is often sufficient, and is usually better than nothing at all. To improve upon this method one would have to search for correlations involving two or more variables, but this gets rather involved. For example, in order to conduct a systematic search for two-variable correlations, that is for integrand extremums caused by two variables r_i and r_j together satisfying some condition, one would have to make $n(n-1)/2$ two-dimensional scatter plots of the extremal coordinates, one plot for each distinct r_ir_j pair. This is obviously much more complicated than making n one-dimensional histograms of the extremal coordinates, which is essentially what we do in the one-variable importance sampling procedure. A correlation between r_i and r_j causing extremal integrand values will show up as a clustering of the extremal coordinate points along some narrow <u>band</u> in the r_ir_j scatter plot, just as a clustering in some narrow <u>interval</u> of the r_i histogram would indicate an extremal producing region of the r_i-axis in the one-variable approach. If such a band is discovered in the r_ir_j scatter plot, one can try to find some two-variable probability density function $P_{ij}(r_i,r_j)$ which is peaked along this band; one then generates r_i and r_j randomly according to $P_{ij}(r_i,r_j)$ instead of uniformly, and divides the integrand by $P_{ij}(r_i,r_j)$, in exact analogy with the one-variable procedure. Alternatively, one might try to find some transformation of variables $(r_i,r_j)\rightarrow(\rho_i,\rho_j)$ which transforms the extremal-producing correlation into one involving ρ_i <u>only</u>, and one could then apply the one-variable importance sampling procedure to ρ_i.

As a simple example of a two-variable correlation, suppose it is found that the integrand $h(r_1,\ldots,r_n)$ assumes extremal values whenever $r_1 \simeq r_2$. This would show up as a clustering of points around the line $r_1 = r_2$ in the r_1r_2 scatter plot of the extremal value coordinates; in fact, if one were sufficiently observant one could probably spot this particular correlation in the coordinate listings of the one-variable importance sampling procedure. One way to deal with this correlation would be to construct a probability density function $P(r_1,r_2)$ which is normalized on and non-zero in the r_1r_2 unit square, and which is peaked along the line $r_1 = r_2$; one then would generate r_1 and r_2 randomly according to $P(r_1,r_2)$, instead of uniformlv, and divide the integrand by $P(r_1,r_2)$. However, a simpler method in this case might be to introduce the change of variables $(r_1,r_2) \rightarrow (\rho_1,\rho_2)$ defined by

$$\left.\begin{aligned} \rho_1 &= \frac{1}{2} + \frac{1}{2}(r_1 - r_2) \\ \rho_2 &= \frac{1}{2}(r_1 + r_2) \end{aligned}\right\} \qquad (4.56)$$

It is easily verified that this transformation carries the r_1r_2 unit square into the $\rho_1\rho_2$ unit square, and that $\partial(\rho_1,\rho_2)/\partial(r_1,r_2) = 1/2$; hence we can simply replace dr_1dr_2 in the integral (4.41) by $2d\rho_1d\rho_2$. We then proceed to generate ρ_1 and ρ_2 uniformly in the unit interval, with r_1 and r_2 determined by inverting the above formulae. The point here is that the $r_1 \simeq r_2$ correlation will now appear as an extremal condition associated with $\rho_1 \simeq 1/2$ <u>independently</u> of ρ_2; this can be easily handled by applying single-variable importance sampling to ρ_1.

Two-variable importance sampling is obviously a much more complicated and demanding enterprise than one-variable importance sampling, but it is nevertheless quite feasible. It does not, however, appear to be feasible at this time to attempt an analogous systematic search for and treatment of correlations involving more than two variables. Indeed, if one-variable importance sampling proves inadequate, then even before attempting two-variable importance sampling one probably should investigate the possibility of using one of the other three variance reducing techniques, perhaps in conjunction with one-variable importance sampling.

Appendix A

PROOF OF THE REJECTION METHOD FOR GENERATING RANDOM NUMBERS

Suppose we are given a set of random numbers $\{x_i'\}$ distributed according to the density function $P_1(x)$, and also a set of random numbers $\{r_i\}$ distributed uniformly in the unit interval. Let $P_2(x)$ be any non-negative integrable (but not necessarily normalized) function, which is bounded by the finite number B_2; more specifically, we require $0 \leqslant P_2(x) \leqslant B_2$ everywhere that $P_1(x)$ is non-vanishing. Suppose we now construct a subset $\{x_i\}$ of the set $\{x_i'\}$ by the following procedure: Draw a random pair x_i' and r_i, and take x_i' to be a member of $\{x_i\}$ if and only if

$$P_2(x_i')/B_2 \geqslant r_i \qquad \text{(A.1)}$$

This process is repeated over and over, using a new pair x_i' and r_i each time, with x_i' being made an element of $\{x_i\}$ whenever (A.1) is found to be satisfied. We shall now prove that the set $\{x_i\}$ constructed in this way is a set of random numbers distributed according to the density function

$$P(x) = \frac{P_1(x) \cdot P_2(x)}{\int_{-\infty}^{\infty} P_1(x') \cdot P_2(x') dx'} \qquad \text{(A.2)}$$

and moreover that the efficiency E of this generating process—i.e., the probability that an arbitrarily chosen element of the set $\{x_i'\}$ will be taken to be an element of the set $\{x_i\}$—is

$$E = \frac{1}{B_2} \int_{-\infty}^{\infty} P_1(x') \cdot P_2(x') dx' \qquad \text{(A.3)}$$

Let $P_0(x)dx$ be the probability that a random draw from the set $\{x_i'\}$, accompanied by a random draw from $\{r_i\}$, will produce an element of the set $\{x_i\}$ which lies between x and x+dx. $P_0(x)dx$ can be expressed in two different ways. From one point of view we can write $P_0(x)dx$ as the product of: (i) the probability that a randomly chosen element x_i' will become a member of the set $\{x_i\}$, times (ii) the probability that a member of the set $\{x_i\}$ will lie between x and x+dx. By definition, the probability (i) is E and the probability (ii) is P(x)dx; hence,

$$P_0(x)dx = [E]\times[P(x)dx] \tag{A.4}$$

From another point of view, we can write $P_0(x)dx$ as the product of: (iii) the probability that a randomly chosen element x_i' will lie between x and x+dx, times (iv) the probability that an x_i'-value which lies between x and x+dx will be accepted as a member of the set $\{x_i\}$. By definition the probability (iii) is $P_1(x)dx$. To find an expression for the probability (iv), we note that it is just the probability that an x_i'-value which lies between x and x+dx will satisfy the acceptance criterion (A.1); in other words, (iv) is the probability that a randomly chosen element r_i will be less than the quantity $P_2(x)/B_2$ (which by hypothesis lies between zero and unity). This probability is precisely $P_2(x)/B_2$, since the probability for an element from $\{r_i\}$ to be less than r, for $0\leqslant r\leqslant 1$, is just $F(r)=r$ [cf. (2.8b)]. Hence, our second expression for $P_0(x)dx$ is

$$P_0(x)dx = [P_1(x)dx] \times [P_2(x)/B_2] \tag{A.5}$$

Now, since P(x) is by definition a properly normalized density function, we have from (A.4)

$$\int_{-\infty}^{\infty} P_0(x)dx = E\int_{-\infty}^{\infty} P(x)dx = E$$

But from (A.5) we also have

$$\int_{-\infty}^{\infty} P_0(x)dx = \frac{1}{B_2}\int_{-\infty}^{\infty} P_1(x)\cdot P_2(x)dx$$

Combining the last two equations yields at once the expression for E in (A.3). Now, (A.4) implies that

$$P(x) = \frac{P_0(x)}{E}$$

Evaluating the numerator from (A.5) and the denominator from (A.3) (which has just been established) gives

$$P(x) = \frac{(1/B_2)P_1(x)\cdot P_2(x)}{(1/B_2)\int_{-\infty}^{\infty} P_1(x')\cdot P_2(x')dx'}$$

thus establishing (A.2). QED.

The "rejection method" presented in Sec. 2-3 is now obtained as a special case of the above procedure: For if we take $\{x_i'\}$ to be uniformly distributed over the finite interval $a \leqslant x \leqslant b$, so that

$$P_1(x) = \begin{cases} 1/(b-a), & \text{for } a \leqslant x \leqslant b \\ 0 \quad , & \text{otherwise} \end{cases} \qquad (A.6)$$

then according to (A.2) the set $\{x_i\}$ will be distributed according to the density function

$$P(x) = \begin{cases} P_2(x)\Big/\int_a^b P_2(x')dx', & \text{for } a \leqslant x \leqslant b \\ 0 & , \text{otherwise} \end{cases} \tag{A.7}$$

and according to (A.3) the generating efficiency will be

$$E = \frac{\int_a^b P_2(x')dx'}{B_2(b-a)} \tag{A.8}$$

Thus, the set $\{x_i\}$ is distributed over the interval $a \leqslant x \leqslant b$ according to $C \cdot P_2(x)$, where C is a normalization constant, and the fraction of the x_i''s which are accepted as elements of the set $\{x_i\}$ is just the ratio of the area under $P_2(x)$ between a and b, to the area under the enclosing box of height B_2 between a and b.

Appendix B

THE JACOBIAN

Consider the transformation T from xyz-space to uvw-space, defined by

$$T: \begin{cases} u = U(x,y,z) \\ v = V(x,y,z) \\ w = W(x,y,z) \end{cases} \qquad \text{(B.1)}$$

We assume that the inverse transformation T^{-1} exists, so that the equations in (B.1) can in principle be "solved" for x,y and z in terms of u,v and w:

$$T^{-1}: \begin{cases} x = X(u,v,w) \\ y = Y(u,v,w) \\ z = Z(u,v,w) \end{cases} \qquad \text{(B.2)}$$

Let $\hat{x},\hat{y},\hat{z}$ and $\hat{u},\hat{v},\hat{w}$ be the orthogonal unit vectors in the two spaces (see Fig. 10). Let P=(x,y,z) be any point in xyz-space, and let P'=(u,v,w) be the image of P under T in uvw-space. Let $d\tau$ be the differential (cubic) volume element in xyz-space built upon the three vectors $\hat{x}dx,\hat{y}dy,\hat{z}dz$ emanating from P, and let $d\tau'$ be the image of $d\tau$ under T in uvw-space. We wish to find out how the volume $d\tau'$ compares with the volume $d\tau(=dxdydz)$. For this we must first find the images $\vec{\varepsilon}_x,\vec{\varepsilon}_y,\vec{\varepsilon}_z$ of the respective vectors $\hat{x}dx,\hat{y}dy,\hat{z}dz$ under T; we can then calculate $d\tau'$ as the volume of the parallelipiped built upon $\vec{\varepsilon}_x,\vec{\varepsilon}_y$ and $\vec{\varepsilon}_z$.

Let Q be the point

$$Q = (x+dx,y,z) = (x,y,z) + (dx,0,0) = P + \hat{x}dx.$$

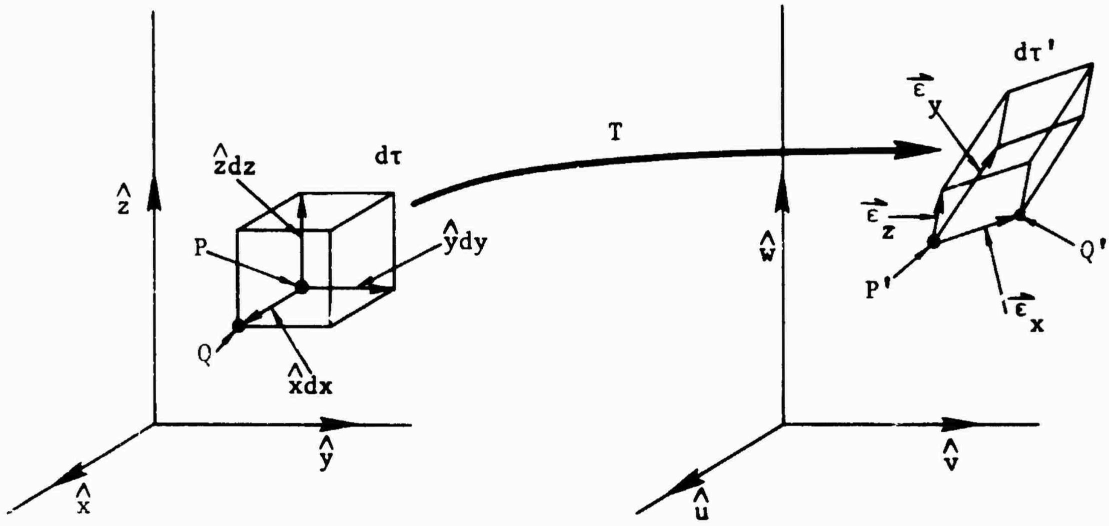

FIGURE 10. Deformation of the differential cubic volume element dxdydz under a hypothetical transformation T from xyz-space to uvw-space.

The image of Q under T is evidently

$$Q' = (u+\frac{\partial u}{\partial x}dx,\ v+\frac{\partial v}{\partial x}dx,\ w+\frac{\partial w}{\partial x}dx)$$

$$= (u,v,w) + (\frac{\partial u}{\partial x}dx, \frac{\partial v}{\partial x}dx, \frac{\partial w}{\partial x}dx)$$

$$= P' + [\hat{u}\frac{\partial u}{\partial x}dx + \hat{v}\frac{\partial v}{\partial x}dx + \hat{w}\frac{\partial w}{\partial x}dx]$$

But since $Q' = P' + \vec{\varepsilon}_x$ (see Fig. 10), then we may conclude that the image $\vec{\varepsilon}_x$ of $\hat{x}dx$ under T is

$$\vec{\varepsilon}_x = \hat{u}\frac{\partial u}{\partial x}dx + \hat{v}\frac{\partial v}{\partial x}dx + \hat{w}\frac{\partial w}{\partial x}dx \qquad \text{(B.3a)}$$

It is of course understood that all the partial derivatives here are evaluated via (B.1) at the point P. In the same way we find

$$\vec{\varepsilon}_y = \hat{u}\frac{\partial u}{\partial y}dy + \hat{v}\frac{\partial v}{\partial y}dy + \hat{w}\frac{\partial w}{\partial y}dy \qquad \text{(B.3b)}$$

and

$$\vec{\varepsilon}_z = \hat{u}\frac{\partial u}{\partial z}dz + \hat{v}\frac{\partial v}{\partial z}dz + \hat{w}\frac{\partial w}{\partial z}dz \qquad \text{(B.3c)}$$

Now, the volume of the parallelipiped built upon any three vectors emanating from the same point is just the absolute value of the so-called "triple scalar product" of these vectors. Thus, we calculate $d\tau'$ as

$$d\tau' = |\vec{\varepsilon}_x \cdot (\vec{\varepsilon}_y \times \vec{\varepsilon}_z)| \qquad \text{(B.4)}$$

The triple scalar product can be written in terms of the orthogonal components of the three vectors in the form of a determinant:

$$\vec{\varepsilon}_x \cdot (\vec{\varepsilon}_y \times \vec{\varepsilon}_z) = \begin{vmatrix} (\vec{\varepsilon}_x)_u & (\vec{\varepsilon}_x)_v & (\vec{\varepsilon}_x)_w \\ (\vec{\varepsilon}_y)_u & (\vec{\varepsilon}_y)_v & (\vec{\varepsilon}_y)_w \\ (\vec{\varepsilon}_z)_u & (\vec{\varepsilon}_z)_v & (\vec{\varepsilon}_z)_w \end{vmatrix}$$

Inserting the specific components from (B.3) we find

$$\vec{\varepsilon}_x \cdot (\vec{\varepsilon}_y \times \vec{\varepsilon}_z) = \frac{\partial(u,v,w)}{\partial(x,y,z)} dxdydz \qquad (B.5)$$

where we have defined the Jacobian of the transformation (B.1) by

$$\frac{\partial(u,v,w)}{\partial(x,y,z)} = \begin{vmatrix} \frac{\partial u}{\partial x} & \frac{\partial v}{\partial x} & \frac{\partial w}{\partial x} \\ \frac{\partial u}{\partial y} & \frac{\partial v}{\partial y} & \frac{\partial w}{\partial y} \\ \frac{\partial u}{\partial z} & \frac{\partial v}{\partial z} & \frac{\partial w}{\partial z} \end{vmatrix} \qquad (B.6)$$

Again we note that all the partial derivatives here are evaluated via (B.1) at the point P. Putting (B.5) into (B.4), and noting that $dxdydz = d\tau$, we conclude that

$$d\tau' = \left| \frac{\partial(u,v,w)}{\partial(x,y,z)} \right| d\tau \qquad (B.7)$$

According to (B.7), the absolute value of the Jacobian (B.6) gives the local volume expansion $d\tau'/d\tau$ produced by the transformation T from xyz-space to uvw-space. By the same token, we may assert that the absolute value of

$$\frac{\partial(x,y,z)}{\partial(u,v,w)} \equiv \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial y}{\partial u} & \frac{\partial z}{\partial u} \\ \frac{\partial x}{\partial v} & \frac{\partial y}{\partial v} & \frac{\partial z}{\partial v} \\ \frac{\partial x}{\partial w} & \frac{\partial y}{\partial w} & \frac{\partial z}{\partial w} \end{vmatrix}$$

in which all of the partial derivatives are evaluated via (B.2) at the point P', gives the local volume expansion accompanying the inverse transformation T^{-1}. Since the net local expansion involved in the successive transformations $(x,y,z)\xrightarrow{T}(u,v,w)\xrightarrow{T^{-1}}(x,y,z)$ must obviously be unity, we have

$$\left|\frac{\partial(u,v,w)}{\partial(x,y,z)}\right| \times \left|\frac{\partial(x,y,z)}{\partial(u,v,w)}\right| = 1$$

whence,

$$\left|\frac{\partial(x,y,z)}{\partial(u,v,w)}\right| = 1\Big/\left|\frac{\partial(u,v,w)}{\partial(x,y,z)}\right| \qquad \text{(B.8)}$$

Suppose now we have an integral of the form

$$I = \iiint_{R'} f(u,v,w)\,du\,dv\,dw$$

where f is some function defined in some region R' of uvw-space.

Heuristically, this integral can be thought of in terms of a partition of uvw-space into infinitesimal volume elements, with the value of the integral being the number obtained by first multiplying the value of f in each infinitesimal volume element by the size of that volume element, and then summing over all elements inside R'. Now, if R' is the image under T of a region R in xyz-space, or equivalently if R is the image of R' under T^{-1}, then one way of partitioning R' would be to proceed as follows: Partition R into infinitesimal volume elements $dxdydz=d\tau$, and then take the volume elements in R' to be the corresponding images $d\tau'$ of the elements $d\tau$ under T. Then we would have

$$\iiint_{R'} f(u,v,w)dudvdw = \iiint_{R'} f(u,v,w)d\tau'$$

where, on the right, u,v and w denote the location of $d\tau'$. Thus, using (B.7) and the fact that $d\tau=dxdydz$, we conclude

$$\iiint_{R'} f(u,v,w)dudvdw = \iiint_{R} f(U(x,y,z),V(x,y,z),W(x,y,z))\left|\frac{\partial(u,v,w)}{\partial(x,y,z)}\right|dxdydz \qquad (B.9)$$

This equation shows how the integral on the left "transforms" under the transformation T^{-1} from uvw-space to xyz-space; it is the rule for "changing integration variables" in a multi-dimensional integral.

It should be apparent from the above results, (B.7) and (B.9), that the Jacobian $\partial(u,v,w)/\partial(x,y,z)$ of the transformation T in (B.1) is the three-dimensional analogue of the derivative du/dx of the one-dimensional transformation $u=U(x)$. Indeed, the definition (B.6) shows that $\partial(u,v,w)/\partial(x,y,z)$ automatically reduces to du/dx in the one-dimensional case.

Appendix C

ADDING INDEPENDENT RANDOM NUMBERS

Let $\{z_{1,i}\}$ and $\{z_{2,i}\}$ be two sets of random real numbers with density functions $P_1(z)$ and $P_2(z)$ respectively. The mean m_k and variance σ_k^2 of the set $\{z_{k,i}\}$ $(k=1,2)$ are then given by [cf. (3.9) and (3.10)]

$$\left.\begin{aligned} m_k &= \int z\, P_k(z)\, dz \\ \sigma_k^2 &= \int (z-m_k)^2 P_k(z) dz = \int z^2 P_k(z)\, dz - m_k^2 \end{aligned}\right\} \qquad \text{(C.1)}$$

Suppose we construct a new set of random numbers $\{Z_i\}$ by drawing a random number from each of the two given sets and forming their sum,

$$Z_i = z_{1,i} + z_{2,i} \qquad \text{(C.2)}$$

Assuming that the draws from the two sets are statistically independent, in that the probability for obtaining any value for $z_{2,i}$ depends only on P_2 and not on the value obtained for $z_{1,i}$, then the density function $P(Z)$ of the new set $\{Z_i\}$ is determined by the following statement: The probability for Z_i to fall in the interval dZ about Z is equal to the product of {the probability for $z_{1,i}$ to fall in the interval dz_1 about z_1} times {the probability for $z_{2,i}$ to fall in the interval $d(Z-z_1)$ about $(Z-z_1)$}, summed over all values of z_1. In mathematical terms we therefore have

$$P(Z)dZ = \int_{z_1} P_1(z_1)dz_1 \cdot P_2(Z-z_1)d(Z-z_1)$$

$$= \int dz_1 P_1(z_1)P_2(Z-z_1)dZ$$

whence

$$P(Z) = \int dz_1 P_1(z_1)P_2(Z-z_1) \qquad (C.3)$$

We may thus compute the mean M and variance Σ^2 of the set $\{Z_i\}$ as follows:

$$M = \int ZP(Z)dZ = \int dZ[Z-z_1+z_1]\int dz_1 P_1(z_1)P_2(Z-z_1)$$

$$= \int d(Z-z_1)\int dz_1[(Z-z_1)+z_1]P_1(z_1)P_2(Z-z_1)$$

$$M = \int dz_2 \int dz_1[z_2+z_1]P_1(z_1)P_2(z_2) \qquad (C.4)$$

$$\Sigma^2 = \int Z^2 P(Z)dZ - M^2 = \int dZ[Z-z_1+z_1]^2 \int dz_1 P_1(z_1)P_2(Z-z_1) - M^2$$

$$= \int d(Z-z_1)\int dz_1[(Z-z_1)+z_1]^2 P_1(z_1)P_2(Z-z_1) - M^2$$

$$\Sigma^2 = \int dz_2 \int dz_1[z_2+z_1]^2 P_1(z_1)P_2(z_2) - M^2 \qquad (C.5)$$

Recognizing that the z_1 and z_2 integrations in (C.4) and (C.5) are independent of each other, and moreover that $\int dz_1 P_1(z_1) = \int dz_2 P_2(z_2) = 1$, it is a simple matter to carry out these integrations using the definitions in (C.1). The results are

$$M = m_1 + m_2 \qquad (C.6)$$

$$\Sigma^2 = \sigma_1^2 + \sigma_2^2 \qquad (C.7)$$

Thus, provided the elements $z_{1,i}$ and $z_{2,i}$ are drawn independently of one another, the mean and variance of the set $\{Z_i\}=\{z_{1,i}+z_{2,i}\}$ will be the sums of the respective means and variances of the sets $\{z_{1,i}\}$ and $\{z_{2,i}\}$. Notice in particular that it is the variances σ_k^2, not the standard deviations σ_k, that add; this has the consequence that Σ is always somewhat less than $\sigma_1+\sigma_2$.

By adding elements from a third set $\{z_{3,i}\}$ to $\{Z_i\}$, we see that the set of summed elements will have mean $(m_1+m_2)+m_3$ and variance $(\sigma_1^2+\sigma_2^2)+\sigma_3^2$. In general, if each element of the set $\{Z_i\}$ is obtained by summing N independently drawn elements from each of the N sets $\{z_{1,i}\}$, $\{z_{2,i}\},\ldots,\{z_{N,i}\}$, then the mean M and variance Σ^2 of the set $\{Z_i\}$ will be

$$M = m_1 + m_2 + \ldots + m_N \tag{C.8}$$

and

$$\Sigma^2 = \sigma_1^2 + \sigma_2^2 + \ldots + \sigma_N^2 \tag{C.9}$$

where m_k and σ_k^2 are the mean and variance respectively of the set $\{z_{k,i}\}$.

In particular, suppose that the N sets of random numbers $\{z_{1,i}\}$, $\{z_{2,i}\},\ldots,\{z_{N,i}\}$ all have the same density function, or in other words, suppose they are all the same set $\{z_i\}$ with mean m and variance σ^2:

$$\{z_{k,i}\} = \{z_i\}; \quad m_k = m, \quad \sigma_k^2 = \sigma^2 \qquad (k=1,2,..,N) \tag{C.10}$$

In this case, each element of the set $\{Z_i\}$ can be regarded as the sum of N independently drawn elements from the same set of random numbers $\{z_i\}$. According to (C.8) and (C.9), the mean M and variance Σ^2 of the set $\{Z_i\}$ will then be given by

$$M = m + m + \ldots + m = Nm \qquad (C.11)$$

$$\Sigma^2 = \sigma^2 + \sigma^2 + \ldots + \sigma^2 = N\sigma^2 \qquad (C.12)$$

Define now the new set of random numbers $\{\tilde{Z}_i\}$ by the rule

$$\tilde{Z}_i \equiv Z_i/N \qquad (C.13)$$

It is easy to show that the mean $\tilde{M}$ and rms deviation $\tilde{\Sigma}$ of the set $\{\tilde{Z}_i\}$ is just going to be 1/N times the mean M and rms deviation Σ, respectively, of the set $\{Z_i\}$. Thus, in view of (C.11) and (C.12) we get

$$\tilde{M} = M/N = (Nm)/N = m \qquad (C.14)$$

and

$$\tilde{\Sigma} = \Sigma/N = (N\sigma^2)^{1/2}/N = \sigma/N^{1/2} \qquad (C.15)$$

We can interpret these results in the following way: If we let $\tilde{Z}$ be the "average" of N independently drawn elements from the set $\{z_i\}$,

$$\tilde{Z} = (z_1 + z_2 + \ldots + z_N)/N, \qquad (C.16)$$

then $\tilde{Z}$ can be regarded as an element of a set of random numbers whose mean is equal to the mean m of the set $\{z_i\}$, and whose rms deviation is equal to the rms deviation σ of the set $\{z_i\}$ divided by $\sqrt{N}$. Thus, we have proved a "weak form" of the Central Limit Theorem, which is discussed in the text in connection with equations (3.11) and (3.12). The Central Limit Theorem further asserts that in the limit of large N the set of random numbers $\{\tilde{Z}_i\}$ becomes a Gaussian distribution; in that case we can assign the numerical confidence limits in (3.12) [cf. also the discussion following (3.19)], which evidently permit a much more quantitative interpretation of the rms deviation $\sigma/\sqrt{N}$ than can be obtained from the development given here.

Appendix D

THE COVARIANCE

Let $f_1(\vec{x})$ and $f_2(\vec{x})$ be two functions defined on a region Ω, and let $f(\vec{x})$ denote their sum:

$$f(\vec{x}) = f_1(\vec{x}) + f_2(\vec{x}) \qquad \text{(D.1)}$$

The variances (or mean square deviations) of these functions with respect to a uniform distribution in Ω are given by

$$\left.\begin{aligned} &\sigma^2 \equiv \mathrm{var}(f{:}P_\Omega) = \langle f^2{:}P_\Omega\rangle - \langle f{:}P_\Omega\rangle^2 \\ \text{and} \\ &\sigma_i^2 \equiv \mathrm{var}(f_i{:}P_\Omega) = \langle f_i^2{:}P_\Omega\rangle - \langle f_i{:}P_\Omega\rangle^2, \quad i=1,2 \end{aligned}\right\} \qquad \text{(D.2)}$$

Here, the bracket $\langle h{:}P_\Omega\rangle$ is defined for any function $h(\vec{x})$ by

$$\langle h{:}P_\Omega\rangle = \int_\infty h(\vec{x})P_\Omega(\vec{x})d\vec{x} = |\Omega|^{-1}\int_\Omega h(\vec{x})d\vec{x} \qquad \text{(D.3)}$$

where $P_\Omega(\vec{x})$ is the density function (3.4) defining the set of random points distributed uniformly over Ω.

We seek a relation between the variance of $f(\vec{x})$ and the variances of $f_1(\vec{x})$ and $f_2(\vec{x})$. By straightforward calculation utilizing (D.2) and (D.3) we have

$$\mathrm{var}(f{:}P_\Omega) = \langle f^2{:}P_\Omega\rangle - \langle f{:}P_\Omega\rangle^2$$

$$= \langle (f_1^2 + 2f_1f_2 + f_2^2){:}P_\Omega\rangle - \langle (f_1 + f_2){:}P_\Omega\rangle^2$$

$$= \langle f_1^2 : P_\Omega \rangle + 2\langle f_1 f_2 : P_\Omega \rangle + \langle f_2^2 : P_\Omega \rangle - (\langle f_1 : P_\Omega \rangle + \langle f_2 : P_\Omega \rangle)^2$$

$$= \langle f_1^2 : P_\Omega \rangle - \langle f_1 : P_\Omega \rangle^2 + \langle f_2^2 : P_\Omega \rangle - \langle f_2 : P_\Omega \rangle^2$$

$$+ 2\langle f_1 f_2 : P_\Omega \rangle - 2\langle f_1 : P_\Omega \rangle \langle f_2 : P_\Omega \rangle$$

Defining, then, the "covariance of $f_1(\vec{x})$ and $f_2(\vec{x})$ with respect to a uniform distribution inside Ω" by

$$\mathrm{cov}(f_1, f_2 : P_\Omega) \equiv \langle f_1 f_2 : P_\Omega \rangle - \langle f_1 : P_\Omega \rangle \langle f_2 : P_\Omega \rangle \tag{D.4}$$

we have the result

$$\mathrm{var}(f : P_\Omega) = \mathrm{var}(f_1 : P_\Omega) + \mathrm{var}(f_2 : P_\Omega) + 2\mathrm{cov}(f_1, f_2 : P_\Omega) \tag{D.5}$$

In fact, it is easy to see that the foregoing arguments admit the slightly more general result

$$\mathrm{var}(a_1 f_1 + a_2 f_2 : P_\Omega) = a_1^2 \mathrm{var}(f_1 : P_\Omega) + a_2^2 \mathrm{var}(f_2 : P_\Omega)$$

$$+ 2a_1 a_2 \mathrm{cov}(f_1, f_2 : P_\Omega) \tag{D.6}$$

where a_1 and a_2 are any two constants.

According to (D.4), the covariance of $f_1(\vec{x})$ and $f_2(\vec{x})$ is just the average of the product of these functions minus the product of their averages. Comparing (D.4) with (D.2) we see that the covariance of any function $h(\vec{x})$ with itself is just its variance:

$$\mathrm{cov}(h,h:P_\Omega) \equiv \mathrm{var}(h:P_\Omega) \tag{D.7}$$

The variance of a function is never negative, as can be seen by writing it in the form [cf. (3.10)]

$$\begin{aligned}\mathrm{var}(h:P_\Omega) &= \Big\langle (h - \langle h:P_\Omega\rangle)^2 : P_\Omega \Big\rangle \\ &= |\Omega|^{-1}\int_\Omega (h(\vec{x}) - \langle h:P_\Omega\rangle)^2 d\vec{x} \geqslant 0\end{aligned} \tag{D.8}$$

However, the covariance of two <u>different</u> functions can be either positive or negative. To get some idea of the limits on the covariance, we introduce the function

$$f_3(\vec{x}) \equiv \mathrm{cov}(f_1,f_2:P_\Omega)f_1(\vec{x}) - \mathrm{cov}(f_1,f_1:P_\Omega)f_2(\vec{x})$$

By using only the definitions of the variance and covariance, together with their implied consequences (D.7) and (D.8), it is a straightforward but slightly tedious calculation to show that the statement $\mathrm{var}(f_3:P_\Omega) \geqslant 0$ implies the inequality

$$|\mathrm{cov}(f_1,f_2:P_\Omega)| \leqslant \sqrt{\mathrm{var}(f_1:P_\Omega)}\cdot\sqrt{\mathrm{var}(f_2:P_\Omega)} \equiv \sigma_1\sigma_2 \tag{D.9}$$

where σ_i is of course the rms deviation of $f_i(\vec{x})$ with respect to a uniform distribution inside Ω [cf. (D.2)].

Combining (D.9) and (D.5) yields the result

$$|\sigma_1 - \sigma_2| \leqslant \sigma \leqslant \sigma_1 + \sigma_2 \tag{D.10}$$

Thus, the rms deviation of $f_1(\vec{x})+f_2(\vec{x})$ assumes its maximum value of $\sigma_1+\sigma_2$ when the covariance of $f_1(\vec{x})$ and $f_2(\vec{x})$ assumes its maximum value of $+\sigma_1\sigma_2$. Similarly, the rms deviation of $f_1(\vec{x})$ and $f_2(\vec{x})$ assumes its minimum value of $|\sigma_1-\sigma_2|$ when the covariance of $f_1(\vec{x})$ and $f_2(\vec{x})$ assumes its minimum value of $-\sigma_1\sigma_2$. If the covariance of $f_1(\vec{x})$ and $f_2(\vec{x})$ happens to vanish, then the rms deviation of $f_1(\vec{x})+f_2(\vec{x})$ will be $\sqrt{\sigma_1^2+\sigma_2^2}$.

It is sometimes convenient to define the "correlation coefficient" ρ of $f_1(\vec{x})$ and $f_2(\vec{x})$ by

$$\rho \equiv \frac{\mathrm{cov}(f_1,f_2:P_\Omega)}{\sqrt{\mathrm{var}(f_1:P_\Omega)}\sqrt{\mathrm{var}(f_2:P_\Omega)}} \tag{D.11}$$

The inequality (D.9) implies that

$$-1 \leqslant \rho \leqslant +1 \tag{D.12}$$

Appendix E

THE VARIANCE OF A FUNCTION OVER A PARTITIONED REGION

Let $f(\vec{x})$ be a function defined in a region Ω, and let Ω be partioned into n subregions $\Omega_1, \Omega_2, \ldots, \Omega_n$. We define α_j to be the ratio of the volume of Ω_j to the volume of Ω:

$$\alpha_j \equiv |\Omega_j| / |\Omega| \tag{E.1}$$

The condition that the union of the n non-overlapping subregions be equal to Ω implies that

$$\sum_{j=1}^{n} \alpha_j = 1 \tag{E.2}$$

We have [cf. (3.5)]

$$|\Omega| \langle f:P_\Omega \rangle = \int_\Omega f(\vec{x})d\vec{x} = \sum_{j=1}^{n} \int_{\Omega_j} f(\vec{x})d\vec{x} = \sum_{j=1}^{n} |\Omega_j| \langle f:P_{\Omega_j} \rangle$$

or

$$\langle f:P_\Omega \rangle = \sum_{j=1}^{n} \alpha_j \langle f:P_{\Omega_j} \rangle \tag{E.3}$$

This expresses the <u>average</u> of $f(\vec{x})$ over Ω in terms of the averages of $f(\vec{x})$ over the various subregions $\Omega_1, \Omega_2, \ldots, \Omega_n$. What we would like to do now is derive an analogous expression for the <u>variance</u> of $f(\vec{x})$ over Ω.

Replacing f by f^2 in (E.3), we have

$$\langle f^2 : P_\Omega \rangle = \sum_{j=1}^{n} \alpha_j \langle f^2 : P_{\Omega_j} \rangle$$

Inserting this and (E.3) into the usual expression for $\mathrm{var}(f:P_\Omega)$, we have

$$\begin{aligned}
\mathrm{var}(f:P_\Omega) &= \langle f^2 : P_\Omega \rangle - \langle f : P_\Omega \rangle^2 \\
&= \sum_j \alpha_j \langle f^2 : P_{\Omega_j} \rangle - \Big(\sum_j \alpha_j \langle f : P_{\Omega_j} \rangle \Big)^2 \\
&= \sum_j \alpha_j \langle f^2 : P_{\Omega_j} \rangle - \sum_{jk}\sum \alpha_j \alpha_k \langle f : P_{\Omega_j} \rangle \langle f : P_{\Omega_k} \rangle \\
&= \sum_j \alpha_j \langle f^2 : P_{\Omega_j} \rangle - \sum_j \alpha_j^2 \langle f : P_{\Omega_j} \rangle^2 \\
&\qquad - \sum_{j \neq k}\sum \alpha_j \alpha_k \langle f : P_{\Omega_j} \rangle \langle f : P_{\Omega_k} \rangle \\
&= \sum_j \alpha_j \Big(\langle f^2 : P_{\Omega_j} \rangle - \langle f : P_{\Omega_j} \rangle^2 \Big) + \sum_j (\alpha_j - \alpha_j^2) \langle f : P_{\Omega_j} \rangle^2 \\
&\qquad - \sum_{j \neq k}\sum \alpha_j \alpha_k \langle f : P_{\Omega_j} \rangle \langle f : P_{\Omega_k} \rangle \\
&= \sum_j \alpha_j \mathrm{var}(f : P_{\Omega_j}) + \sum_j \alpha_j (1 - \alpha_j) \langle f : P_{\Omega_j} \rangle^2 \\
&\qquad - 2 \sum_{j<k}\sum \alpha_j \alpha_k \langle f : P_{\Omega_j} \rangle \langle f : P_{\Omega_k} \rangle
\end{aligned}$$

Using (E.2) we can rewrite the second term in the following way:

$$\sum_j \alpha_j(1 - \alpha_j)\langle f:P_{\Omega_j}\rangle^2 = \sum_j \alpha_j\Big(\sum_{k\neq j}\alpha_k\Big)\langle f:P_{\Omega_j}\rangle^2$$

$$= \sum\sum_{j\neq k} \alpha_j\alpha_k\langle f:P_{\Omega_j}\rangle^2$$

$$= \sum\sum_{j<k} \alpha_j\alpha_k\langle f:P_{\Omega_j}\rangle^2 + \sum\sum_{j<k} \alpha_j\alpha_k\langle f:P_{\Omega_k}\rangle^2$$

$$= \sum\sum_{j<k} \alpha_j\alpha_k(\langle f:P_{\Omega_j}\rangle^2 + \langle f:P_{\Omega_k}\rangle^2)$$

Inserting this into the previous expression for $\mathrm{var}(f:P_\Omega)$, we find

$$\mathrm{var}(f:P_\Omega) = \sum_j \alpha_j \mathrm{var}(f:P_{\Omega_j}) + \sum\sum_{j<k} \alpha_j\alpha_k$$

$$\times\ (\langle f:P_{\Omega_j}\rangle^2 + \langle f:P_{\Omega_k}\rangle^2 - 2\langle f:P_{\Omega_j}\rangle\langle f:P_{\Omega_k}\rangle)$$

or equivalently

$$\mathrm{var}(f:P_\Omega) = \sum_{j=1}^{n} \alpha_j \mathrm{var}(f:P_{\Omega_j})$$

$$+ \sum_{j=1}^{n}\sum_{\substack{k=1 \\ j<k}}^{n} \alpha_j\alpha_k\Big(\langle f:P_{\Omega_j}\rangle - \langle f:P_{\Omega_k}\rangle\Big)^2 \qquad \text{(E.4)}$$

The result (E.4) expresses $\mathrm{var}(f:P_\Omega)$ as a sum of two terms: the first term is due to the variations in $f(\vec{x})$ within the various subregions, and the second term is due to the variations in $f(\vec{x})$ among the various sub-egions. In particular, since the second term is never negative we have the corollary

$$\mathrm{var}(f:P_{\Omega}) \geqslant \sum_{j=1}^{n} \alpha_j \mathrm{var}(f:P_{\Omega_j}) \tag{E.5}$$

which is to be contrasted with the result (E.3).

Appendix F

OPTIMUM APPORTIONMENT IN STRATIFIED SAMPLING

In the "stratified sampling" procedure discussed in Sec. 4-4, the region of integration Ω is partitioned into n subregions $\Omega_1, \Omega_2, \ldots, \Omega_n$, and a separate Monte Carlo integration is performed in each subregion. The square of the uncertainty associated with this Monte Carlo procedure is [cf. (4.24)]

$$\Delta^{*2} = \sum_{j=1}^{n} |\Omega_j|^2 \sigma_j^2 N_j^{-1} \tag{F.1}$$

where σ_j^2 is the variance of the integrand with respect to the uniform distribution of points inside Ω_j. We wish to find the set of values $\tilde{N}_1, \tilde{N}_2, \ldots, \tilde{N}_n$ which minimizes (F.1), subject to the condition

$$\sum_{j=1}^{n} N_j = N \tag{F.2}$$

Suppose that, for a given set of N_j values, we vary each by a small amount δN_j. These variations are presumed to be consistent with condition (F.2) but otherwise quite arbitrary; in other words, all we require of the small variations $\delta N_1, \delta N_2, \ldots, \delta N_n$ is that they be such that

$$\delta N = \delta \sum_{j=1}^{n} N_j = 0$$

or equivalently,

$$\sum_{j=1}^{n} \delta N_j = 0 \tag{F.3}$$

Now these small variations in the N_j's will induce a small variation in Δ^{*2} in the amount

$$\begin{aligned}\delta\Delta^{*2} &= \delta \sum_{j=1}^{n} |\Omega_j|^2 \sigma_j^2 N_j^{-1} \\ &= \sum_{j=1}^{n} |\Omega_j|^2 \sigma_j^2 \delta N_j^{-1} \\ &= \sum_{j=1}^{n} |\Omega_j|^2 \sigma_j^2 (-N_j^{-2} \delta N_j)\end{aligned}$$

$$\delta\Delta^{*2} = -\sum_{j=1}^{n} |\Omega_j|^2 \sigma_j^2 N_j^{-2} \delta N_j \tag{F.4}$$

In particular, if the variations δN_j are all taken about the minimizing values $\tilde{N}_j$, then the variation in Δ^{*2} will evidently vanish; hence, we have

$$\sum_{j=1}^{n} |\Omega_j|^2 \sigma_j^2 \tilde{N}_j^{-2} \delta N_j = 0 \tag{F.5}$$

Now, the only way for (F.5) to hold for every set of variations $(\delta N_1, \delta N_2, \ldots, \delta N_n)$ which satisfies (F.3) is for the quantity multiplying δN_j in (F.5) to be a constant, independent of j:

$$|\Omega_j|^2 \sigma_j^2 \tilde{N}_j^{-2} \equiv C^2, \quad j=1,\ldots,N \tag{F.6}$$

For then and only then will (F.5) be satisfied without requiring of the variations δN_j anything more than is required by conditions (F.3). We therefore have

$$\tilde{N}_j = C^{-1}|\Omega_j|\sigma_j \tag{F.7}$$

which says that $\tilde{N}_j$ is proportional to the product of the volume $|\Omega_j|$ times the rms deviation σ_j.

The constant C is determined simply by requiring the values $\tilde{N}_j$ to satisfy (F.2):

$$\sum_{i=1}^{n} \tilde{N}_i = C^{-1} \sum_{i=1}^{n} |\Omega_i|\sigma_i = N \tag{F.8}$$

Therefore,

$$\tilde{N}_j = \left(N\Big/\sum_{i=1}^{n} |\Omega_i|\sigma_i\right)|\Omega_j|\sigma_j \tag{F.9}$$

With this result we can immediately calculate the minimum value of Δ^*:

$$\begin{aligned}
\Delta^{*2}_{\min} &= \sum_{j=1}^{n} |\Omega_j|^2\sigma_j^2\tilde{N}_j^{-1} \\
&= \left(\sum_{i=1}^{n} |\Omega_i|\sigma_i/N\right) \sum_{j=1}^{n} |\Omega_j|^2\sigma_j^2|\Omega_j|^{-1}\sigma_j^{-1} \\
&= N^{-1}\left(\sum_{i=1}^{n} |\Omega_i|\sigma_i\right) \sum_{j=1}^{n} |\Omega_j|\sigma_j
\end{aligned}$$

whence

$$\Delta^*_{\min} = N^{-1/2} \sum_{j=1}^{n} |\Omega_j| \sigma_j \tag{F.10}$$

Appendix G

OPTIMUM DENSITY FUNCTION FOR IMPORTANCE SAMPLING

The importance sampling procedure discussed in Sec. 4-5 is based on the fact that the integral

$$I = \int_{\Omega} f(\vec{x})d\vec{x} \tag{G.1}$$

can be regarded as the mean of the function $f(\vec{x})/P(\vec{x})$, taken with respect to the set of random points $\{\vec{x}_i\}$ distributed over Ω according to the density function $P(\vec{x})$:

$$I = \int_{\Omega} [f(\vec{x})/P(\vec{x})]P(\vec{x})d\vec{x} = \langle f/P:P \rangle \tag{G.2}$$

If we <u>estimate</u> this mean by averaging a finite number N of randomly chosen elements of the set $\{f(\vec{x}_i)/P(\vec{x}_i)\}$, then the square of the uncertainty in our estimate will be

$$\Delta^{*2} = \frac{\langle (f/P)^2:P \rangle - \langle f/P:P \rangle^2}{N}$$

$$= N^{-1}\left| \int_{\Omega} [f(\vec{x})/P(\vec{x})]^2 P(\vec{x})d\vec{x} - I^2 \right|$$

or

$$\Delta^{*2} = N^{-1}\{J - I^2\} \tag{G.3}$$

where we have defined

$$J \equiv \int_{\Omega} f^2(\vec{x})P^{-1}(\vec{x})d\vec{x} \tag{G.4}$$

We now ask the question, what density function $P(\vec{x})$ will render Δ^{*2} a minimum? Since the quantities f, Ω and N are presumed fixed, then clearly Δ^{*2} will be a minimum if and only if J is a minimum.

Let $P_0(\vec{x})$ be the minimizing function; that is, $P_0(\vec{x})$ satisfies the conditions

$$P_0(\vec{x}) \geqslant 0 \quad \text{for} \quad \vec{x}\varepsilon\Omega \tag{G.5}$$

and

$$\int_{\Omega} P_0(\vec{x})d\vec{x} = 1 \tag{G.6}$$

and furthermore, of <u>all</u> functions $P(\vec{x})$ which satisfy these conditions, $P_0(\vec{x})$ causes J in (G.4) to assume the smallest value. Form the family of functions $P_{\varepsilon}(\vec{x})$ according to

$$P_{\varepsilon}(\vec{x}) = P_0(\vec{x}) + \varepsilon\eta(\vec{x}) \tag{G.7}$$

where ε is a real variable (the family parameter), and where $\eta(\vec{x})$ is <u>any</u> function which satisfies the condition

$$\int_{\Omega} \eta(\vec{x})d\vec{x} = 0 \tag{G.8}$$

With (G.6), this condition on $\eta(\vec{x})$ evidently insures that each function $P_{\varepsilon}(\vec{x})$ in (G.7) satisfies the requirement

$$\int_{\Omega} P_{\varepsilon}(\vec{x})d\vec{x} = 1 \tag{G.9}$$

It follows, then, that the quantity

$$J(\varepsilon) \equiv \int_{\Omega} f^2(\vec{x})P_{\varepsilon}^{-1}(\vec{x})d\vec{x} \qquad (G.10)$$

has a minimum at $\varepsilon=0$. Differentiating (G.10) with respect to ε gives

$$J'(\varepsilon) = \frac{d}{d\varepsilon}\int_{\Omega} f^2(\vec{x})P_{\varepsilon}^{-1}(\vec{x})d\vec{x}$$

$$= \int_{\Omega} f^2(\vec{x})\frac{\partial}{\partial\varepsilon}P_{\varepsilon}^{-1}(\vec{x})d\vec{x}$$

$$= \int_{\Omega} f^2(\vec{x})[-P_{\varepsilon}^{-2}(\vec{x})\ \frac{\partial}{\partial\varepsilon}P_{\varepsilon}(\vec{x})]d\vec{x}$$

or, with (G.7),

$$J'(\varepsilon) = -\int_{\Omega} f^2(\vec{x})P_{\varepsilon}^{-2}(\vec{x})\eta(\vec{x})d\vec{x} \qquad (G.11)$$

The fact that $J(\varepsilon)$ has a minimum at $\varepsilon=0$ implies that $J'(0)=0$; hence, we have

$$\int_{\Omega} f^2(\vec{x})P_0^{-2}(\vec{x})\eta(\vec{x})d\vec{x} = 0 \qquad (G.12)$$

Now, the only way for (G.12) to hold for every function $\eta(\vec{x})$ which satisfies (G.8) is for the quantity multiplying $\eta(\vec{x})$ in (G.12) to be constant over Ω:

$$f^2(\vec{x})P_0^{-2}(\vec{x}) \equiv C^2, \quad \text{all } \vec{x}\varepsilon\Omega \qquad (G.13)$$

For then and only then will (G.12) be satisfied without requiring of $\eta(\vec{x})$ anything more than is required by condition (G.8). Combining (G.13) with (G.5) yields the result

$$P_0(\vec{x}) = C^{-1}|f(\vec{x})|, \quad C>0 \tag{G.14}$$

The precise value of C is then determined through requirement (G.6):

$$1 = \int_\Omega P_0(\vec{x}')d\vec{x}' = C^{-1}\int_\Omega |f(\vec{x}')|d\vec{x}' \tag{G.15}$$

Therefore, the density function $P_0(\vec{x})$ which minimizes J, and hence Δ^{*2}, is

$$P_0(\vec{x}) = |f(\vec{x})| \Big/ \int_\Omega |f(\vec{x}')|d\vec{x}' \tag{G.16}$$

With this result we can easily calculate the minimum value of J:

$$\begin{aligned} J_{min} &= J(\varepsilon=0) = \int_\Omega f^2(\vec{x})P_0^{-1}(\vec{x})d\vec{x} \\ &= \int_\Omega |f(\vec{x}')|d\vec{x}' \cdot \int_\Omega f^2(\vec{x})|f(\vec{x})|^{-1}d\vec{x} \\ &= \int_\Omega |f(\vec{x}')|d\vec{x}' \cdot \int_\Omega |f(\vec{x})|d\vec{x} \end{aligned}$$

whence

$$J_{min} = \left(\int_\Omega |f(\vec{x})|d\vec{x}\right)^2 \tag{G.17}$$

The minimum value of Δ^{*2} is thus [cf. (G.3)]

$$\Delta^{*2}_{\min} = N^{-1}\left|J_{\min} - I^2\right|$$

or

$$\Delta^{*2}_{\min} = N^{-1}\left|\left(\int_{\Omega}|f(\vec{x})|d\vec{x}\right)^2 - \left(\int_{\Omega} f(\vec{x})d\vec{x}\right)^2\right| \qquad \text{(G.18a)}$$

Using the identity $a^2-b^2=(a-b)(a+b)$, this result can also be written in the form

$$\Delta^{*2}_{\min} = 4N^{-1}\int_{\Omega^-}|f(\vec{x})|d\vec{x}\cdot\int_{\Omega^+}|f(\vec{x})|d\vec{x} \qquad \text{(G.18b)}$$

where Ω^- is the subregion of Ω in which $f(\vec{x})<0$ and Ω^+ is the subregion of Ω in which $f(\vec{x})>0$. It follows from (G.18b) that $\Delta^{*2}_{\min}$ will <u>vanish</u> if and only if $f(\vec{x})$ never changes sign inside Ω (in which case either $|\Omega^-|$ or $|\Omega^+|$ vanishes).

Appendix H

SOME ONE-VARIABLE IMPORTANCE SAMPLING DENSITY FUNCTIONS

The basic idea behind the one-variable importance sampling technique described in Sec. 4-6 may be briefly stated as follows: In a Monte Carlo calculation of the integral

$$I = \int_0^1 dr_1 \cdots \int_0^1 dr_n h(r_1, \ldots, r_n) \tag{H.1}$$

we may generate any particular coordinate r_j inside the unit interval according to a non-uniform density function $P(r_j)$, provided we take $h(r_1, \ldots, r_n)/P(r_j)$ as the quantity to be averaged instead of $h(r_1, \ldots, r_n)$. Doing this can be advantageous if $h(r_1, \ldots, r_n)$ happens to assume extremal values whenever the variable r_j falls inside some small subinterval (α_j, β_j) of the unit interval. For then, by choosing $P(r_j)$ to be large whenever $\alpha_j < r_j < \beta_j$, the denominator in the quantity being averaged will moderate the extremal behavior of the numerator in that critical r_j interval. As a result, the variance of the values being averaged will be reduced, and the Monte Carlo uncertainty will be made smaller.

In this appendix we shall describe several simple density functions $P(r)$ which can be used for one-variable importance sampling. Generally, any such density function must satisfy the requirements

$$P(r) > 0 \quad \text{for } 0 \leqslant r \leqslant 1 \tag{H.2a}$$

$$\int_0^1 P(r)dr = 1 \tag{H.2b}$$

The condition (H.2a), that P(r) not vanish inside the unit interval, is simply to insure boundedness of $h(r_1,\ldots,r_n)/P(r)$ [r stands of course for one of the r_i variables], and can be relaxed at any point where $h \to 0$ faster than $P \to 0$. As discussed in Sec. 4-6, in order to actually importance sample the variable r according to a given density function P(r), it is most convenient to have a computer subroutine which does the following:

1° Accepts as input a number r' in the unit interval. [Normally, r' will be a given random number from the uniform distribution in the unit interval.]

2° Calculates and outputs that value r which satisfies

$$r' = F(r) \tag{H.3}$$

where F is the distribution function corresponding to the density function P. [This number r will be used in evaluating $h(r_1,\ldots,r_n)$.]

3° Calculates and outputs the value P(r) for the r value found in 2°. [This number P(r) will be divided into $h(r_1,\ldots,r_n)$.]

In what follows we shall develop equations from which such a computer subroutine can be written for three simple functions P(r) that the author has found to be particularly useful. In addition, we shall see how one can go about constructing a one-variable importance sampling subroutine for an arbitrarily shaped density function P(r).

▶ $P(r) \propto e^{-\Gamma r} \quad (\Gamma>0)$

This density function is often useful in cases where the integrand h assumes extremal values whenever r is near 0. The size of the parameter Γ is to be chosen, normally by trial-and-error, to be commensurate with the degree of peaking of h near r=0; the greater this peaking the larger Γ should be. The normalization condition (H.2b) allows us to determine the normalization constant, and one easily finds that the correctly normalized density function is

$$P(r) = \frac{\Gamma}{1 - e^{-\Gamma}}\, e^{-\Gamma r} \qquad (0\leqslant r\leqslant 1) \tag{H.4}$$

The calculation of the corresponding distribution function is straightforward, and yields

$$F(r) \equiv \int_0^r P(r')dr' = \frac{1 - e^{-\Gamma r}}{1 - e^{-\Gamma}} \tag{H.5}$$

We note as a check that, as r increases from 0 to 1, F(r) also increases from 0 to 1. Putting (H.5) into (H.3) and solving for r yields

$$r = -\frac{1}{\Gamma}\ell n[1 - r'(1-e^{-\Gamma})] \tag{H.6}$$

Thus, given any r' between 0 and 1 [item 1°], we calculate r from (H.6) [item 2°]; then, using this value of r, we calculate P(r) from

(H.4) [item 3°].

► $P(r) \propto e^{-\Gamma(1-r)}$ $(\Gamma>0)$

This density function is often useful in cases where the integrand h assumes extremal values whenever r is near 1. The larger these extremal values are, the greater the value of Γ should be. Since this density function is just the mirror image of (H.4) about the line r=1/2, the normalization constant is the same:

$$P(r) = \frac{\Gamma}{1 - e^{-\Gamma}} e^{-\Gamma(1-r)} \qquad (0 \leqslant r \leqslant 1) \tag{H.7}$$

The calculation of the corresponding distribution function is straightforward, and yields

$$F(r) \equiv \int_0^r P(r')dr' = \frac{e^{\Gamma r} - 1}{e^{\Gamma} - 1} \tag{H.8}$$

Putting this into (H.3) and solving for r yields

$$r = \frac{1}{\Gamma} \ell n[1 + r'(e^{\Gamma}-1)] \tag{H.9}$$

Thus, given any r' between 0 and 1 [item 1°], we calculate r from (H.9) [item 2°]; then, using this value of r, we calculate P(r) from (H.7) [item 3°].

► $P(r) \propto 1/[(r - r_0)^2 + \Gamma^2]$ $(0 \leqslant r_0 \leqslant 1,\ \Gamma>0)$

This density function is often useful in cases where the integrand is peaked whenever

$$\mathrm{Max}(0,r_0-\Gamma) < r < \mathrm{Min}(1,r_0+\Gamma)$$

where r_0 is any point in the unit interval. The greater the integrand peaking at r_0, the smaller the value of Γ should be [in contrast to the two previous density functions]. The normalization constant is easily determined by requiring $P(r)$ to satisfy (H.2b), and the correctly normalized density function is found to be

$$P(r) = \left(\frac{\Gamma}{A+B}\right)\frac{1}{(r - r_0)^2 + \Gamma^2}, \quad 0\leqslant r\leqslant 1 \tag{H.10}$$

where the constants A and B are defined by

$$A \equiv \arctan\left[\frac{1 - r_0}{\Gamma}\right], \quad B \equiv \arctan\left[\frac{r_0}{\Gamma}\right] \tag{H.11}$$

The calculation of the corresponding distribution function yields

$$F(r) \equiv \int_0^r P(r')dr' = \left(\frac{1}{A+B}\right)\left(\arctan\left[\frac{r - r_0}{\Gamma}\right] + B\right) \tag{H.12}$$

Putting this into (H.3) and solving for r yields

$$r = r_0 + \Gamma\tan[r'(A + B) - B] \tag{H.13}$$

Thus, given any r' between 0 and 1 [item 1°], we calculate the corresponding value of r from (H.13) [item 2°], where A and B are defined in (H.11); then, using this value of r, we calculate the value P(r) from (H.10)[item 3°].

The sharpness of the peaking of each of the foregoing three density functions is controlled by the single parameter Γ, which parameter can be varied in the "preliminary runs" to determine a more or less optimal value [see Sec. 4-6]. The shape of each of these density functions is clearly restricted by its analytical form; however, it usually turns out that one's knowledge of the dependence of the function $h(r_1,\ldots,r_n)$ on any one of its variables is so meager that one usually cannot take advantage of very much flexibility in the shape of importance sampling density functions. Nevertheless, situations do occasionally arise in which one clearly sees the need to importance sample some variable r according to a density function P(r) of a very specific shape. If, in such a case, it appears to be impractical to find an analytic form for P(r) which is simple enough that its distribution function can be calculated and inverted, then one can always <u>approximate</u> the desired P(r) curve as closely as necessary by a <u>piecewise</u> <u>linear</u> <u>curve</u>, as indicated in Fig. 11. The point here is that it is fairly easy to write a computer subroutine which will: (<u>i</u>) accept the "pivot points" (ρ_i,σ_i) on input data cards; (<u>ii</u>) scale the ordinates σ_i so that the total area under the piecewise linear curve is unity, thereby rendering the piecewise linear curve a properly normalized density function; and (<u>iii</u>) calculate, for any given value r' between 0 and 1, that value r for which the area under the piecewise linear curve between the vertical lines through 0 and r is equal to r'. This last step is of course equivalent to inverting the distribution function corresponding to the piecewise linear density function.

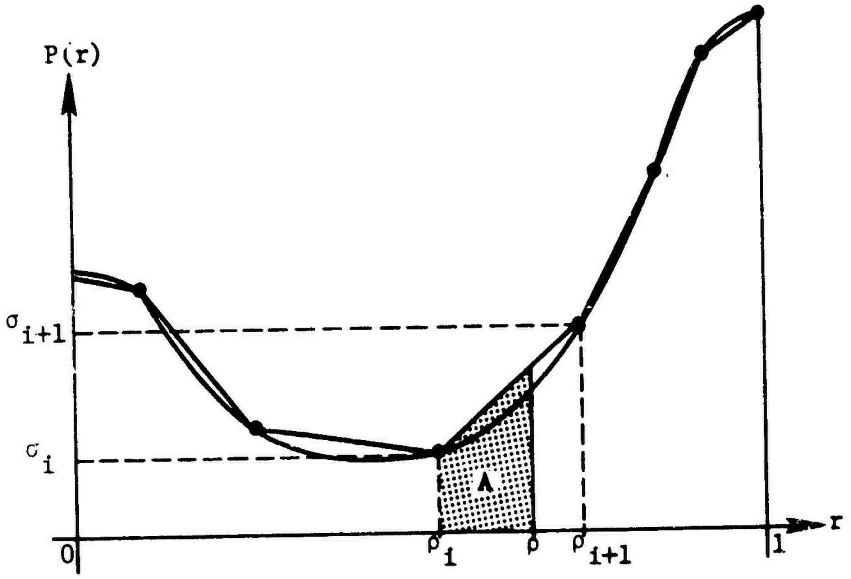

FIGURE 11. Approximating an arbitrarily shaped, one-variable density function by a piecewise linear density function.

To carry out item (<u>ii</u>) above, one need only recognize that the "raw area" A_i under the trapezoid between $r=\rho_i$ and $r=\rho_{i+1}$ is [see Fig. 11]

$$A_i = (\rho_{i+1} - \rho_i)\cdot\sigma_i + \frac{1}{2}\cdot(\rho_{i+1} - \rho_i)\cdot(\sigma_{i+1} - \sigma_i) \qquad \text{(H.14)}$$

and the total area under the piecewise linear curve can be made unity if each ordinate σ_j is simply divided by the total raw area $\Sigma_i A_i$. Having thus normalized the piecewise linear curve, one can carry out item (<u>iii</u>) by observing that the area under the curve between $r=\rho_i$ and $r=\rho$ ($\rho_i \leqslant \rho \leqslant \rho_{i+1}$) is [cf (H.14)]

$$A = \alpha(\rho_{i+1} - \rho_i)\cdot\sigma_i + \frac{1}{2}\cdot\alpha(\rho_{i+1} - \rho_i)\cdot\alpha(\sigma_{i+1} - \sigma_i)$$

where α is the fractional distance of ρ from ρ_i to ρ_{i+1}:

$$\alpha \equiv (\rho - \rho_i)/(\rho_{i+1} - \rho_i)$$

From these two equations one can show that the value of ρ which corresponds to a <u>given</u> value of A ($A \leqslant A_i$) is

$$\rho = \begin{cases} \rho_i + A/\sigma_i & , \text{ if } \sigma_{i+1} = \sigma_i \\ \rho_i + \left(\sqrt{\sigma_i^2 + 2Am_i} - \sigma_i\right)\Big/m_i, & \text{ if } \sigma_{i+1} \neq \sigma_i \end{cases} \qquad \text{(H.15)}$$

where

$$m_i \equiv (\sigma_{i+1} - \sigma_i)/(\rho_{i+1} - \rho_i) \qquad \text{(H.16)}$$

In applying these equations it is important to note that it is <u>not</u>

necessary for the points ρ_i to be equally spaced along the r-axis; all that is required is that each ρ_i be greater than ρ_{i-1}, and each σ_i be positive.

With the ability to generate random numbers in the unit interval according to <u>any</u> bounded, piecewise linear density function, we obviously have great flexibility for carrying out the one-variable importance sampling procedure described in Sec. 4-6. However, as mentioned previously, one's knowledge of the behavior of the integrand $h(r_1,\ldots,r_n)$ as a function of any of the variables r_i is usually so limited that one usually cannot take full advantage of this flexibility. In practice, therefore, a simple analytic density function, like one of the three described earlier, usually proves adequate.

Appendix I

REMARKS ON THE MARKOV CHAIN MONTE CARLO METHOD

In the field of statistical mechanics one is confronted with the task of calculating macroscopic properties of systems composed of very many identical microscopic components which interact according to known (or hypothesized) laws. Examples are the calculation of the equation of state of a gas composed of molecules which interact via a specified intermolecular force, and the calculation of the magnetization of a lattice of atoms whose magnetic moments interact with each other and with an external magnetic field according to the laws of electrodynamics. Typically, these calculations require the evaluation of one or more n-dimensional integrals, where n is of the order of the number of microscopic components (e.g., molecules or atoms) in the system under study. In most cases these integrals cannot be calculated either by analytical methods or by classical numerical methods; in fact, it usually happens that even the conventional Monte Carlo method breaks down for these problems.[†] As a consequence, workers in this field usually employ another Monte Carlo method, one which was first used by Metropolis and co-workers in 1953 to calculate

[†] However, in Sec. 2-10 we derive a set of generating formulae [see (2.94)] which might be used to calculate by the conventional Monte Carlo procedure the equilibrium properties of a one-dimensional gas of impenetrable molecules.

the equation of state of a two-dimensional gas of hard disks (Ref. 10). The distinguishing feature of this Monte Carlo approach is that it utilizes the mathematical concept of a "Markov chain" or a "Markovian random walk". It is not our purpose in this report to lay out the theory of this special Monte Carlo technique; however, the successes of the Markov chain approach have become so well known that any introduction to Monte Carlo methods would be incomplete without some discussion of it. Therefore, in this appendix we shall give a very brief description of the Markov chain Monte Carlo method; for more explicit discussions, the reader should consult Chapter 9 of Hammersley and Handscomb (Ref. 1), and the extensive review article by Wood (Ref. 11) and references contained therein.

The general problem is once again to evaluate an integral of the form

$$I = \int_{\Omega} f(\vec{x})P(\vec{x})d\vec{x} \qquad (I.1)$$

where $P(\vec{x})$ is a probability density function normalized o.:r the finite, n-dimensional region Ω. In the context of statistical mechanical applications, the multi-dimensional variable $\vec{x}$ usually specifies the physical "state" of some system (e.g., the spatial coordinates of all the molecules in a gas, or the magnetic moment orientations of all the atoms in a magnetic substance); Ω is the set of all physically allowable states; $f(\vec{x})$ denotes the value of some dynamical variable f when the system is in state $\vec{x}$; and $P(\vec{x})d\vec{x}$ denotes the probability for a randomly chosen svstem from an appropriate "statistical ensemble" of systems to be in some state in the infinitesimal region $d\vec{x}$ around $\vec{x}$. The form of the function $P(\vec{x})$ is given by the laws of statistical mechanics, and is usually taken to be the "microcanonical ensemble" probability density

$$P(\vec{x}) = e^{-U(\vec{x})/kT} \Big/ \int_\Omega e^{-U(\vec{x}')/kT} d\vec{x}' \qquad (\vec{x}\epsilon\Omega)$$

where $U(\vec{x})$ is the total energy of the system in the state $\vec{x}$, T the absolute temperature, and k Boltzmann's constant. In this context, the value of I in (I.1) would be interpreted physically as "the equilibrium value of f at temperature T". The dimensionality n of Ω is usually quite large, and $P(\vec{x})$ is usually analytically complicated and exceedingly small over most of Ω.[†] More often than not, these conditions combine to make it totally impractical to generate random points according to $P(\vec{x})$ by either the inversion method or the rejection method; as a result, the Markov chain Monte Carlo method frequently offers the only hope for evaluating I.

In the Markov chain approach, it is convenient to regard the

[†] With reference to the problem of the one-dimensional gas of impenetrable rods discussed in Sec. 2-10, the region Ω considered here would be more properly associated with the region Σ in (2.95) rather than the region Ω in (2.77). The complicated nature of $P(\vec{x})$ in that case arises from the fact that $P(\vec{x})$ vanishes everywhere inside Σ (2.95) except in that extremely small and oddly shaped subregion Ω (2.77).

space of the variable $\vec{x}$ as being *discrete* rather than continuous. For this we can imagine setting up in the space of $\vec{x}$ a very fine n-dimensional cubic grid or mesh, which subdivides Ω into a total of B "cells" of equal size $\Delta\vec{x}$. We number these cells by the index i in any convenient fashion, and we let $\vec{x}_i$ denote the center of the i^{th} cell. Thus, whereas originally the system could be in a non-denumerably infinite number of states $\vec{x}\varepsilon\Omega$, we now suppose that the system can only be in one of the B states $\vec{x}_1,\vec{x}_2,\ldots,\vec{x}_B$ in Ω. It is of no concern to us, either theoretically or practically, how extremely large B is, just so long as it is finite.[†] Now, since $P(\vec{x})$ in (I.1) is normalized over Ω, then provided the mesh size $\Delta\vec{x}$ is taken sufficiently small we can write with negligible error

$$I \equiv \frac{\int_{\Omega} f(\vec{x})P(\vec{x})d\vec{x}}{\int_{\Omega} P(\vec{x})d\vec{x}} = \frac{\sum_{i=1}^{B} f(\vec{x}_i)P(\vec{x}_i)\Delta\vec{x}}{\sum_{j=1}^{B} P(\vec{x}_j)\Delta\vec{x}}$$

Hence,

[†] This situation is always realized on a digital computer: If 10^{-m} is the smallest number the computer can handle without "underflowing", then the computer will necessarily regard the n-dimensional unit cube as a collection of $B=10^{mn}$ discrete points on an n-dimensional cubic lattice.

$$I = \sum_{i=1}^{B} f(\vec{x}_i)\pi_i \tag{I.2}$$

where

$$\pi_i \equiv P(\vec{x}_i)\Big/\sum_{j=1}^{B} P(\vec{x}_j) \quad (i=1,\ldots,B) \tag{I.3}$$

is the (correctly normalized) probability associated with the discrete state $\vec{x}_i$.

If we could somehow generate N random states $\vec{x}_{i_1}, \vec{x}_{i_2}, \ldots, \vec{x}_{i_N}$ according to the probabilities π_i in (I.3), then we could approximate I as the average of f over these states,

$$I \simeq \frac{1}{N}\sum_{j=1}^{N} f(\vec{x}_{i_j}) \tag{I.4}$$

with the associated uncertainty being determined in the usual way from the variance of the values being averaged. However, by hypothesis it is not possible to generate such a set of random states by any of the conventional Monte Carlo methods.

At this point we introduce the (seemingly unrelated) concept of a Markovian random walk over the lattice of points $\vec{x}_i$ inside Ω. More specifically, we consider a walk over these discrete states which is governed solely by a set of one-step probabilities p_{ij} $(i,j=1,\ldots B)$, defined by

$p_{ij} \equiv$ the probability that, if the walker is at state $\vec{x}_i$, the next step will carry the walker to state $\vec{x}_j$. (I.5)

It should be clearly understood that *such a random walk over the states $\vec{x}_i$ has nothing at all to do with the actual time behavior of the physical system under study*; the random walk is merely a mathematical artifice which we are introducing in order to effect a calculation of the integral I. The adjective "Markovian" simply means that the probability of walking from $\vec{x}_i$ to $\vec{x}_j$ is independent of the past history of the walk (i.e., of where the walker was before coming to state $\vec{x}_i$); if the situation were otherwise, the random walk would be "non-Markovian", and would not be describable simply by the B^2 probabilities in (I.5).

Since p_{ij} is a probability, we have

$$0 \leqslant p_{ij} \leqslant 1 \quad (i,j=1....,B) \tag{I.6}$$

Furthermore, since the walker will always step from $\vec{x}_i$ to *some* $\vec{x}_j$, we have

$$\sum_{j=1}^{B} p_{ij} = 1 \quad (i=1,...,B) \tag{I.7}$$

The probability $p_{ij}^{(2)}$ that the walker will go from $\vec{x}_i$ to $\vec{x}_j$ in *two* steps is, by the multiplication and addition laws for probabilities,

$$p_{ij}^{(2)} = \sum_{k=1}^{B} p_{ik}p_{kj}$$

In general we define the n-step probabilities $p_{ij}^{(n)}$ (i,j=1,...,B) by

$$p_{ij}^{(n)} \equiv \text{the probability of walking from state } \vec{x}_i \text{ to state } \vec{x}_j \text{ in } n \text{ steps.} \tag{I.8}$$

It is clear that $p_{ij}^{(1)}=p_{ij}$, and that for any $n\geqslant 2$

$$p_{ij}^{(n)} = \sum_{k_1=1}^{B} \sum_{k_2=1}^{B} \cdots \sum_{k_{n-1}=1}^{B} p_{ik_1} p_{k_1k_2} \cdots p_{k_{n-1}j} \tag{I.9}$$

If we regard the probabilities p_{ij} as elements of a B×B matrix $\underset{\sim}{P}$,

$$\underset{\sim}{P} \equiv \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1B} \\ p_{21} & p_{22} & \cdots & p_{2B} \\ \cdot & \cdot & \cdot & \\ \cdot & \cdot & \cdot & \\ p_{B1} & p_{B2} & \cdots & p_{BB} \end{bmatrix} \tag{I.10}$$

then (I.9) simply says that the B×B matrix $\underset{\sim}{P}^{(n)}$ of the n-step probabilities $p_{ij}^{(n)}$ is obtained by multiplying $\underset{\sim}{P}$ by itself n times:

$$\underset{\sim}{P}^{(n)} = \underset{\sim}{P}^{n} \tag{I.11}$$

Granted, then, that we can conceive of a random walk over the discrete states $\vec{x}_i$ in Ω, the walk being characterized by a one-step probability matrix $\underset{\sim}{P}$, what does this have to do with estimating the quantity I in (I.2)? The answer to this question is supplied in part by the following theorem [a rigorous proof of which may be found in Chapter XV of Ref. 8]:

<u>Theorem</u>: Suppose that, for a given set of state probabilities π_i, we have a set of one-step probabilities p_{ij} which satisfy the following two conditions:

A (Ergodic Condition). If $\vec{x}_i$ and $\vec{x}_j$ are any two states in Ω with non-vanishing probabilities π_i and π_j, then there exists some finite n for which $p_{ij}^{(n)} > 0$.

B (Steady State Condition).

$$\sum_{i=1}^{B} \pi_i p_{ij} = \pi_j, \quad \text{for all } j=1,\ldots,B \tag{I.12}$$

Then we will have, independently of i and for all j:

$$\lim_{n\to\infty} p_{ij}^{(n)} = \pi_j \tag{I.13}$$

In other words, if for a given set of state probabilities π_i we can construct a random walk matrix $\underset{\sim}{P}$ which satisfies conditions (A) and (B) above, then by starting a random walk in <u>any</u> state $\vec{x}_i$, the probability of winding up in state $\vec{x}_j$ <u>after a sufficiently large number of steps</u> will be π_j. The ergodic condition (A) essentially requires $\underset{\sim}{P}$ to be such that any possible state $\vec{x}_j$ be reachable from any possible state $\vec{x}_i$ by a finite number of steps. The steady state condition (B) essentially requires $\underset{\sim}{P}$ to be such that if many walkers are randomly distributed over the discrete states $\vec{x}_i$ according to the probabilities π_i, this random distribution of the walkers will not be altered if each walker steps once according to $\underset{\sim}{P}$. From a strictly mathematical point of view, the steady state condition (I.12) requires the one-step probability matrix $\underset{\sim}{P}$ to have the state probability vector

$$\underset{\sim}{\pi} \equiv \begin{bmatrix} \pi_1 \\ \pi_2 \\ \cdot \\ \cdot \\ \cdot \\ \pi_B \end{bmatrix} \tag{I.14}$$

as a left eigenvector with eigenvalue unity; that is, $\underset{\sim}{\pi}\cdot\underset{\sim}{P} = \underset{\sim}{\pi}$.

Before inquiring into the possibility of finding, for a given set of state probabilities π_i, a set of one-step probabilities p_{ij} satisfying conditions (A) and (B), let us consider what we should do once we have found such a set. On the basis of conventional Monte Carlo theory we might proceed as follows: Starting out in any allowable initial state $\vec{x}_{i_0}$, take a "sufficiently large number n" of successive steps according to p_{ij}, and then regard the state arrived at at the n^{th} step to be a state randomly chosen inside Ω according to the prescribed state probabilities π_i. Repeating this process N-1 more times would give us the N random points $\vec{x}_{i_j}$ necessary to calculate the estimate of I in (I.4). An obvious question here is precisely what constitutes a "sufficiently large value" of n; in other words, what is the smallest value of n that will allow us to read (I.13) as $p_{ij}^{(n)} \simeq \pi_j$ for all i and j?

We cannot answer this question in any realistic situation; however, it is somewhat academic anyway since the foregoing procedure is not the one normally followed in the Markov chain Monte Carlo method. Instead, the usual procedure is to approximate I by the so-called chain average

$$I \simeq \frac{1}{N} \sum_{t=1}^{N} f(\vec{x}_i^{(t)}) \qquad (I.15)$$

where $\vec{x}_i^{(t)}$ is the discrete state arrived at at the t^{th} step of an N-step random walk according to $\underset{\sim}{P}$. [The sequence of states $\vec{x}_i^{(1)}$, $\vec{x}_i^{(2)}$, ..., $\vec{x}_i^{(N)}$ is referred to as a "realization of the Markov chain".] The justification for this procedure is the "Central Limit Theorem for Markov chains" [see p. 119 of Ref. 11]; this theorem says in effect that if conditions (A) and (B) cited above are satisfied, then provided N is sufficiently large the chain average on the right side of (I.15) will be Gaussian distributed about the value I with a standard deviation proportional to $1/\sqrt{N}$.

Like the Central Limit Theorem which forms the basis for the conventional Monte Carlo method [cf. Sec. 3-2], the Central Limit Theorem for Markov chains does not tell us how large N must be in order for the Gaussian distribution to be realized. However, in view of the asymptotic nature of (I.13), one has the intuitive suspicion that N will usually have to be very much larger for the Markov chain process than for the ordinary Monte Carlo process in order for Gaussian results to be obtained. If the convergence of (I.13) were "immediate", that is if $p_{ij}^{(n)} \simeq \pi_j$ for n=1, then the Markov chain average in (I.15) would evidently be equivalent to the conventional Monte Carlo average in (I.4). But to the extent that the convergence of (I.13) requires large values of n, it seems reasonable to suppose that the Markov chain average in (I.15) will require many more terms in order to yield results comparable to what might be expected from a conventional Monte Carlo average.

Let us now consider how one might go about constructing, for a given set of state probabilities π_i, a set of one-step probabilities p_{ij} satisfying conditions (A) and (B). It turns out that, so far as the ergodic condition (A) is concerned, one usually can only hope for the best. Certainly, one should not use any walking scheme which is obviously incapable of reaching all possible states $\vec{x}_i$ in Ω. However, difficulties can arise if Ω consists of "islands" of high probability states in a "sea" of very low probability states, in which case the probability of walking from one island to another will, for most walking schemes, be very small. For reasonable values of N one might very well just walk around on only one island, and if the function f varied significantly from island to island erroneous results would obviously be obtained. Unfortunately, one is rarely able to definitely rule out this possibility in any specific calculation.

The steady state condition (B) can usually be satisfied rather easily. Notice first of all that condition (I.12) depends only on the relative magnitudes of the state probabilities π_i; furthermore, it is seen from (I.3) that the probabilities π_i themselves depend only on the relative magnitudes of the density function P at various discrete states. Consequently we can always write π_i simply as $P(\vec{x}_i)$ in any conveniently *un*normalized form, and thereby avoid the (usually impossible) tasks of (*i*) evaluating the normalization constant for the density function $P(\vec{x})$, and (*ii*) evaluating the normalizing denominator in (I.3) for π_i.

One way of setting up a one-step matrix $\underset{\sim}{P}$ which satisfies the steady state condition (B) is as follows: Let p^*_{ij} be any set of one-step probabilities which satisfies, in addition to the usual conditions (I.6) and (I.7), the condition

$$p^*_{ij} \equiv p^*_{ji} \qquad \text{(I.16)}$$

[It is usually easy to set up such a *symmetric* one-step probability matrix $\underset{\sim}{P}^*$ in any specific application.] We then define $\underset{\sim}{P}$ in terms of $\underset{\sim}{P}^*$ and $\underset{\sim}{\pi}$ according to

$$\left.\begin{aligned} (i{\neq}j):\quad p_{ij} &= \begin{cases} p^*_{ij}\cdot\pi_j/\pi_i, & \text{if } \pi_j/\pi_i<1 \\ p^*_{ij} \quad\ \ , & \text{if } \pi_j/\pi_i\geqslant 1 \end{cases} \\ p_{ii} &= p^*_{ii} + \sum_{\substack{j=1 \\ (\pi_j<\pi_i)}}^{B} p^*_{ij}(1-\pi_j/\pi_i) \end{aligned}\right\} \qquad \text{(I.17)}$$

It is not difficult to show that this set p_{ij} satisfies (I.6), (I.7) and (I.12) [see p. 119 of Ref. 1 for a proof]. We might note that the satisfaction of the steady state condition (I.12) essentially results from the fact that (I.17) is so constructed that

$$\pi_i p_{ij} = \pi_j p_{ji} \tag{I.18}$$

From this it follows, using (I.7), that

$$\sum_{i=1}^{B} \pi_i p_{ij} = \sum_{i=1}^{B} \pi_j p_{ji} = \pi_j \sum_{i=1}^{B} p_{ji} = \pi_j$$

which is just (I.12). We should remark that (I.17) is not the only scheme for constructing $\underset{\sim}{P}$ so that (I.18) is satisfied; furthermore, (I.18) itself is not a necessary condition for satisfying the steady state condition (I.12).

To actually realize a Markov chain according to (I.17), suppose the t^{th} state of the Markov chain is $\vec{x}_i$ and we wish to determine the $(t+1)^{th}$ state. First, we pick a <u>tentative</u> state $\vec{x}_j$ according to p^*_{ij}, and we calculate the probability ratio π_j/π_i for these two states.[†] If $\pi_j/\pi_i \geqslant 1$, we take $\vec{x}_j$ to be the $(t+1)^{th}$ state of the Markov chain. If $\pi_j/\pi_i < 1$, we pick a random number r from the uniform distribution in the unit interval, and we compare π_j/π_i with r; if $\pi_j/\pi_i \geqslant r$ we take $\vec{x}_j$ as the $(t+1)^{th}$ state, but if $\pi_j/\pi_i < r$ we take $\vec{x}_i$ as the $(t+1)^{th}$ state. It is important to note that p_{ii} is usually not zero, so one should have no misgivings about winding up with the same state $\vec{x}_i$ for both the t^{th} and the $(t+1)^{th}$ states of the chain; indeed [see p. 122 of Ref. 11],

[†] For the microcanonical ensemble function $P(\vec{x})$ mentioned just after (I.1), the π_j/π_i ratio is simply $\exp(-\Delta U_{ji}/kT)$, where $\Delta U_{ji} = U(\vec{x}_j) - U(\vec{x}_i)$ is the change in the total energy of the system in going from state $\vec{x}_i$ to state $\vec{x}_j$.

one will introduce errors if one *forces* a genuine change of state at every step of the random walk.

For further details on applying the Markov chain Monte Carlo method to specific problems in statistical mechanics, the reader should consult the review article by Wood [Ref. 11] and references contained therein. We shall leave the subjec at this point with the following general comments. The Markov chain Monte Carlo method has shown itself capable of calculating certain kinds of integrals which cannot presently be calculated either by classical methods or by the conventional Monte Carlo method. However, the Markov chain method requires a considerable amount of care and expertise on the part of the user, more so than does the conventional Monte Carlo method. A prime source of uneasiness when using the Markov chain approach is that one is almost never sure to what extent the ergodic condition (A) is satisfied. In addition, in view of the asymptotic nature of the result (I.13), one cannot help but wonder how large N will have to be in order for the Central Limit Theorem for Markov chains to truly govern the accuracy of the approximation (I.15). For these reasons (which this writer freely admits may be due to his own lack of experience with Markov chain calculations) this writer is of the opinion that the Markov chain Monte Carlo method should be attempted only if the conventional Monte Carlo method is clearly inapplicable. Others may disagree with this opinion; perhaps the one-dimensional gas of impenetrable rods considered in Sec. 2-10 might afford a vehicle for comparing the efficiency and reliability of the two Monte Carlo methods.

REFERENCES

1. Hammersley, J. M. and D. C. Handscomb. *Monte Carlo Methods*. London, Methuen, 1964.

2. Fluendy, M. "Monte Carlo Methods," in *Markov Chains and Monte Carlo Calculations in Polymer Science*, ed. by G. Lowry. New York, Marcel Dekker, 1970. Pp. 45-90.

3. Gillespie, D. T. *Ph.D. Dissertation*, Johns Hopkins University, 1968. (University Microfilms, Box 1346, Ann Arbor, Mich., Order No. 68-16,419.)

4. Arnold Engineering Development Center. *Three-Particle Collision Integrals for Thermal Conductivity, Viscosity and Self-Diffusion of a Gas of Hard Spherical Molecules. Part II. Calculations*, by D. T. Gillespie and J. V. Sengers, University of Maryland. Tullahoma, Tenn., AEDC, 1973. 224 pp. (AEDC-TR-73-171.)

5. Chambers, R. P. "Random-Number Generation", *IEEE Spectrum*, Vol. 4, No. 2 (February 1967), pp. 48-56.

6. Knuth, D. E. *The Art of Computer Programming*. Reading, Mass., Addison-Wesley, 1969. Volume 2, pp. 1-160.

7. Boeing Scientific Research Laboratories. *One-Line Random Number Generators and their Use in Combinations* by G. Marsaglia and T. A. Bray. March 1968. 9 pp. (BSRL Document D1-82-0689.)

8. Feller, W. *An Introduction to Probability Theory and Its Applications*, Vol. 1, 3rd ed. New York, John Wiley, 1968.

9. Kahn, H. "Use of Different Monte Carlo Sampling Techniques." in Symposium on Monte Carlo Methods (1954), ed. by H. A. Meyer. New York, John Wiley, 1956. Pp. 146-190.

10. Metropolis, N., A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller. "Equation of State Calculations by Fast Computing Machines," J. Chem. Phys., Vol. 21, No. 6 (June 1953), pp 1087-1092.

11. Wood, W. W. "Monte Carlo Studies of Simple Liquid Models," in Physics of Simple Liquids, ed. by H.N.V. Temperley, J. S. Rowlinson, and G. S. Rushbrooke. Amsterdam, North-Holland, 1968, Pp. 116-230.